



Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

PHD THESIS

TRINITY COLLEGE DUBLIN

THE UNIVERSITY OF DUBLIN

SCHOOL OF COMPUTER SCIENCE AND STATISTICS

Tús Maith: A Framework to Curate Natural Language Entry Points to Address the Initial Exploration Problem in RDF Knowledge Graphs

Author:

Claire McNamara

Supervisor:

Professor Declan O'Sullivan

Co-Supervisor:

Professor Lucy Hederman

Submitted

June 14, 2026

THESIS SUBMITTED TO THE
UNIVERSITY OF DUBLIN, TRINITY COLLEGE
IN FULFILLMENT OF THE REQUIREMENTS OF THE DEGREE OF
DOCTOR OF PHILOSOPHY

Declaration of Work

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

Signed: Claire McNamara

Date: 06/03/2026

Abstract

Knowledge graphs (KGs) provide a powerful means of representing and linking rich complex information, yet their structure, ontologies, and query mechanisms create significant barriers for lay users attempting to begin exploration of them. This thesis investigates the challenge faced by lay users at the moment of first contact with an unfamiliar complex KG and introduces the Initial Exploration Problem (IEP), defined as a pre-goal state characterised by three interrelated barriers: scope uncertainty, ontology opacity, and query incapacity. These barriers prevent users from identification of a meaningful starting point for exploration and expose a limitation in many existing KG exploration interfaces which typically presuppose that the user already knows where to begin.

To address this problem, the thesis introduces Tús Maith, a multistage framework designed to support the creation of curated natural language question and answer (CuQA) entry points into KGs. These CuQAs function as cognitive and linguistic scaffolds that reveal both the scope and structure of a KG while abstracting away from ontology complexity and formal query languages. The framework is built around the lifecycle of the CuQA through three stages: Competency Question (CQ) template derivation from the KG schema, CQ template instantiation using KG entities, and domain expert curation. While CQs are traditionally used as ontology engineering artefacts, they are repurposed in this research as exploratory instruments capable of exposing the scope and structure of a KG through natural language.

Tús Maith is instantiated as TMv1, which incorporates large language models (LLMs) to support scalable generation and instantiation of CQ templates while foregrounding human oversight through a human-in-the-loop curation process. Two technical experiments investigate the capacity of LLMs to derive CQ templates from the KG schema and to instantiate these CQ templates under strict constraints of entity fidelity and hallucination control. Three user evaluations examine the effectiveness of CuQAs as exploration starting points for lay users and the viability of Tús Maith to support domain experts using the Virtual Record Treasury of Ireland (VRTI) KG. Results indicate that CuQAs are associated with positive post-use perceptions of scope revelation and approachability of the KG when compared to a neutral reference point, providing partial support for their ability to mitigate the IEP. Evaluation with domain experts further demonstrates that TMv1 can function as a viable system for curating such entry points. A proof-of-concept application to the LetterSampo correspondence KG provides supporting evidence of the portability of Tús Maith, as instantiated using LLMs becoming TMv1, beyond the VRTI KG across other digital humanities KGs.

This thesis makes one major and two minor contributions. The major contribution is the design, instantiation, implementation, and evaluation of Tús Maith, a framework that implements the lifecycle of CuQA entry points into KGs and demonstrates how such entry points can support early-stage exploration by lay users. The first minor contribution is the

identification and characterisation of the Initial Exploration Problem (IEP), providing a conceptual framework that unifies three barriers (scope uncertainty, ontology opacity, and query incapacity) and offers a vocabulary for analysing first-contact KG exploration and designing interventions targeting this stage of interaction. The second minor contribution is an empirical investigation of LLM-assisted scaffolding for KGs, demonstrating how LLMs can support the scalable generation and instantiation of CQ templates while requiring domain expert oversight.

Together, these contributions advance methods for improving the accessibility of complex KGs and foreground the importance of supporting the earliest stage of user interaction: determining where exploration can begin.

Tús Maith Leath na hOibre - a good start is half the work.

Acknowledgements

I would like to start by thanking my wonderful supervisors: Professor Declan O'Sullivan and Professor Lucy Hederman. It is not hyperbolic to say that completing this PhD would never have been possible without their unwavering support and patience. They believed in me right from the second year of my undergraduate Computer Science degree, when they took me on as a bright-eyed, research-hungry student volunteering during term time. Declan later became my official supervisor during the Laidlaw Research and Leadership Scholarship in the summers of my second and third years, and again for my final year project in the fourth year of my undergraduate degree, continuing with Lucy as my co-supervisor to the end of my PhD. This belief and support over the course of seven and a half years, spanning three years of undergraduate research and four and a half years of this PhD, has enabled the fulfilment of a dream I have held since childhood – to pursue knowledge to the highest level possible.

I would next like to thank my family, my wonderful parents, Caroline and Ken, and my incredible, creative, and extraordinary sister, Aoife. My family have exhibited a great degree of patience in listening to me talk about my research, have been steadfast in their support, and similarly have never once wavered in their belief in my ability to achieve my PhD. My mom Caroline's kindness in creating what can only be described as a "writers' retreat", sustaining me with delicious food and excellent company, has made this thesis-writing journey, dare I say it, enjoyable.

My talented and brilliant partner, Edd, also deserves a special mention. He too has listened to my research for hours on end, proofread emails, heard all of my hopes and fears, and never once did he doubt that I could complete my PhD. His exceptional patience, kindness, and support played a key role in this thesis becoming a reality.

I would further like to thank the people involved with the Virtual Record Treasury of Ireland (VRTI) project for all of their insight and help in supporting and conducting my evaluations. Being even a small part of such a wonderful, engaging, and intelligent group has been such a privilege.

My friends too have, somehow, stayed by my side throughout this PhD journey. I would like to thank them for being there for me even when I was not the most communicative.

Finally, I would like to dedicate this thesis to my cat, Angel. Words cannot do justice to adequately explain how much she means to me and the role she has played in me getting to this point. She has been there through every late night, every difficult moment, every triumph, and has been by my side since I was nine years old giving almost nineteen years of unconditional love. We (Angel and I) would like to also thank the garden birds (specifically: the robins, blackbirds, house sparrows, starlings, ring-collared doves, jackdaws, magpies, rooks, hooded crows, and even the wood pigeon) that have provided endless entertainment out the window as I wrote this thesis.

The research conducted in this thesis was primarily funded by the Irish Research Council under project ID GOIPG/2021/116, partially funded by the School of Computer Science and Statistics Trinity College Dublin, and supported by the Research Ireland ADAPT Centre for Digital Content Technology (grant number 13/RC/2106_P2).

Contents

1	Introduction	1
1.0.1	Motivation	1
1.0.2	Research Questions	6
1.0.3	Research Objectives	7
1.0.4	Research Methodology	7
1.0.5	Evaluations	11
1.0.6	Thesis Contributions	14
1.0.7	Thesis Overview	17
2	Background	18
2.1	Large Language Models (LLMs)	18
2.1.1	Architecture	18
2.2	Ontologies	19
2.2.1	CIDOC Conceptual Reference Model (CIDOC CRM)	19
2.2.2	The Virtual Record Treasury of Ireland (VRTI) KG	20
2.3	Digital Humanities KG User Interfaces (UIs)	21
2.3.1	Existing KG-Based Historical UIs	22
3	The Initial Exploration Problem (IEP) and Related Work	24
3.1	The Theoretical Underpinnings of the Initial Exploration Problem (IEP)	24
3.1.1	Foundations in Information Behaviour: Uncertainty and ASK	25
3.1.2	Frameworks for Process: Exploratory Search and Sensemaking	25
3.1.3	Cognitive Foundations: Load, Foraging, and Scaffolding	26
3.1.4	Where do we Start? Orientation, Onboarding, and Serendipitous Discovery	26
3.1.5	Synthesis	26
3.2	The Manifestation of the Initial Exploration Problem (IEP) in KGs	27
3.2.1	Complex Ontologies and Technicality	27
3.2.2	Scale and Interconnectedness	29
3.2.3	The IEP in the Digital Humanities (DH) Domain	29
3.2.4	The Characteristics of the IEP	29
3.2.5	The IEP Framing as a Unified Concept	30
3.3	The IEP as a Constraint on KG Interface Design	30
3.3.1	Solution Development Pre-Problem Formulation	31

3.3.2	Interaction Primitives and Epistemic Preconditions	31
3.3.3	Paradigms of KG Exploration as Compositions of Primitives	32
3.3.4	The Missing Interaction Primitive: Scope Revelation	33
3.3.5	Implications for IEP-Oriented Interface Design	33
3.4	Summary and Implications	34
4	The Design, Instantiation, and Implementation of Tús Maith	35
4.1	Chapter Objectives	35
4.2	The Tús Maith Framework	36
4.2.1	Design Motivation: The Curated Question and Answer (CuQA) Life-cycle	36
4.2.2	Framework Definition	37
4.2.3	Human Judgement	38
4.2.4	Entity-Centric Query Anchoring Model	39
4.3	TMv1: Instantiating the Tús Maith Framework	40
4.3.1	Rationale for LLM-Supported Stages	40
4.3.2	Oversight and Control Strategy	42
4.3.3	Stage 1: Existing CQ Template Derivation Systems and Approaches that Utilise LLMs	42
4.3.4	Stage 2: LLMs and their use in CQ Template Instantiation	43
4.4	TMv1: System Architecture	44
4.4.1	External Dependencies	45
4.4.2	Component Responsibilities	45
4.4.3	Data and Control Flow	46
4.4.4	Full System Architecture	47
4.5	TMv1: System Specification	48
4.5.1	Stage 1: LLM-supported CQ Template Derivation	48
4.5.2	Stage 2: LLM-supported CQ Template Instantiation	52
4.5.3	Stage 3: Expert Curation	56
4.6	TMv1 Prototype Configuration for User Evaluation 3	57
4.7	Summary	58
5	Evaluation	61
5.1	Evaluation Overview and Rationale	61
5.1.1	Evaluation Goals	61
5.1.2	Mixed-Methods Evaluation Strategy	62
5.1.3	Mapping of Evaluations to Research Objectives	62
5.2	Evaluation Instruments and Analysis Methods	63
5.2.1	User Evaluation Instruments	63
5.2.2	LLM Evaluation Instruments	65
5.2.3	Statistical Procedures Used In Relevant Evaluations	68
5.3	Baseline System and Integration Context	69
5.4	Domain Expert Context	69

5.5	User Evaluation 1: CuQAs as Entry Points	70
5.5.1	Exploratory Research Questions	71
5.5.2	Experimental Setup	71
5.5.3	Evaluation Procedure	73
5.5.4	Data Collection Instruments	73
5.5.5	Results and Design Implications	74
5.5.6	Summary of Key Findings	78
5.6	User Evaluation 2: Addressing the Initial Exploration Problem	78
5.6.1	Hypothesis and Research Question	79
5.6.2	Experimental Setup	79
5.6.3	Statistical Approach Used	84
5.6.4	Results and Implication for CuQAs as Mitigation of the IEP	86
5.6.5	Summary of Key Findings	92
5.7	LLM Experiment 1: TMv1 Stage 1, LLM-supported CQ Template Derivation	93
5.7.1	Hypotheses	93
5.7.2	Experimental Setup	94
5.7.3	Results	99
5.7.4	Comparative Insights	101
5.7.5	Summary of Key Findings	102
5.8	LLM Experiment 2: TMv1 Stage 2, LLM-supported CQ Template Instantiation	102
5.8.1	Hypothesis	103
5.8.2	Problem Definition and Research Question	103
5.8.3	Experiment Methodology	105
5.8.4	Results	113
5.8.5	Discussion	118
5.8.6	Summary of Key Findings	120
5.9	User Evaluation 3: Assessing the Viability of TMv1	121
5.9.1	Hypothesis and Research Questions	122
5.9.2	Experimental Setup	122
5.9.3	Evaluation Procedure	124
5.9.4	Results and Implications	131
5.9.5	Summary of Key Findings	138
5.10	Assessing Tús Maith Framework’s Portability: A Proof-of-Concept with LetterSampo	139
5.10.1	Tús Maith Applied to the LetterSampo KG	139
5.10.2	The Stages of Tús Maith as Applied to LetterSampo using TMv1	140
5.10.3	Validation Step	143
5.10.4	LetterSampo Proof-of-Concept Conclusion	143
5.11	Evaluation Conclusions	145
5.11.1	Cross-Evaluation Synthesis	145

6	Conclusion	148
6.1	Research Objectives	148
6.1.1	Research Objectives	148
6.1.2	Research Contributions	152
6.1.3	Major Contribution: The Tús Maith Framework for Curated Natural Language Exploratory Entry Points into Knowledge Graphs (KGs)	152
6.1.4	Minor Contribution 1: The Identification and Characterisation of the Initial Exploration Problem	153
6.1.5	Minor Contribution 2: Empirical Evaluations of large language model (LLM)-Assisted Scaffolding for KGs	153
6.2	Limitations	154
6.2.1	Limitations of Claims	154
6.2.2	Limitations of TMv1	154
6.3	Future Work	156
6.3.1	Applying TMv1 beyond the VRTI KG and Proof-of-Concept	156
6.3.2	Support for Multi-Hop, Complex CuQA Creation	156
6.3.3	Support for Multilingualism	156
6.4	Impact	157
6.4.1	Research Dissemination	157
6.4.2	Potential Impact	157
6.5	Final Remarks	158
	Bibliography	175
A	Appendix	176
A.1	Recruitment Emails	176
A.1.1	User Evaluation 1: Recruitment Email	176
A.1.2	LLM Experiment: 1 Recruitment Email	176
A.1.3	User Evaluation 2: Recruitment Email	177
A.1.4	User Evaluation 3: Recruitment Email	178
A.2	Questionnaires	180
A.2.1	LLM Experiment 1: Questionnaire	180
A.3	User Evaluation Scripts	184
A.3.1	User Evaluation 3: Script	184
A.4	Ethics Applications	186
A.4.1	User Evaluation 1: Ethics Application	186
A.4.2	User Evaluation 2: Ethics Application	193
A.4.3	LLM Experiment 1: Ethics Application	201
A.4.4	User Evaluation 3: Ethics Application	209
A.5	LLM Prompts Used	219
A.5.1	LLM Experiment 1: Prompt	219
A.5.2	LLM Experiment 2: Prompts Used	220
A.5.3	User Evaluation 3: Prompts Used	221

A.6	Entity Ranking Algorithm	223
A.7	Normalisation Rules	228
A.7.1	Generic Text Normalisation (<code>normalise_text</code>)	228
A.7.2	Answer Normalisation	228
A.7.3	Person Name Normalisation (<code>normalise_name</code>)	229
A.7.4	Tokenisation for Jaccard Similarity (<code>tokenise_set</code>)	230
A.8	SPARQL Queries	230
A.8.1	Stage 1: SPARQL Query, Generate Classes and Properties and Relationships	230
A.8.2	Stage 2a: SPARQL Query, Determine Existence of Entities that Match a Search Term	231
A.8.3	Stage 2b: SPARQL Query, Retrieve the Necessary Information about a Given Entity in Order to Rank it	231
A.8.4	Stage 2c: SPARQL Query, Retrieve the Semantic Neighbourhood of a Given Entity	236
A.8.5	LetterSampo Application: SPARQL Queries	241
B	Future Work Extended	245
B.1	Augmenting TMv1 to Support Multi-Hop CuQA Generation	245
B.1.1	Design Motivation	245
B.1.2	KG Topology Informed CQ Template Generation	246
B.1.3	Multi-Hop CQ Templates as a Subset of Meta-Paths	246
B.1.4	Stage 1 Augmented: Proposed System for Topology-Informed Multi-Hop CQ Template Generation	249
B.1.5	Stage 2 Augmented: Addressing Instantiation Challenges for Multi-Hop CQ Templates	250
B.1.6	Evaluation Methodologies for Complex CuQAs	250
B.1.7	Future Work Conclusion	251

List of Figures

1.1	The Tús Maith framework: A framework to create CuQAs.	8
1.2	TMv1: An LLM-supported instantiation of the Tús Maith framework. . . .	9
1.3	TMv1: A UML activity diagram denoting the process by which CuQAs are generated.	10
2.1	Homepage of the LetterSampo UI.	22
2.2	Homepage of the VRTI Explorer.	23
4.1	The Tús Maith framework: A framework to create CuQAs.	38
4.2	TMv1: An LLM-supported instantiation of the Tús Maith framework. . . .	40
4.3	TMv1 Component Diagram.	48
4.4	Full TMv1 system configuration showing both frozen (precomputed) and live components for User Evaluation 3.	59
4.5	Live components of TMv1 as experienced by domain experts during User Evaluation 3.	60
5.1	UE1 Implementation of Tús Maith.	72
5.2	VRTI Explorer without Tús Maith.	73
5.3	Feature Helpfulness Ranking.	75
5.4	Feature Helpfulness by VRTI Familiarity Level - Most Helpful	76
5.5	Feature Helpfulness by Familiarity Level - Least Helpful	77
5.6	CuQAs implemented within the VRTI Explorer as Featured Questions. . . .	80
5.7	Piechart presenting the number of participants who had/had not used the VRTI Explorer prior to the evaluation (n=35).	87
5.8	Participant's self-reported familiarity with the VRTI project (n=35) where 1="Not familiar at all" and 5="Extremely familiar".	87
5.9	Participants' response to "I felt confident knowing where to start exploring the VRTI dataset" (n=35) where 1="Not confident at all" and 5="Extremely confident".	88
5.10	LLM and GT Construction for LLM Experiment 2.	108
5.11	Stage two of TMv1 with JSON Validation Step.	120
5.12	TMv1: UI implementation demonstrating the information shown after clicking to view provenance.	120
5.13	Conveyance of the use of AI to domain experts.	125
5.14	Stage 1: Choosing of CQ Templates.	125

5.15	Stage 2a: Choosing of search term to generate entities.	126
5.16	Stage 2b: Choosing of two person and two place related entities to fill in the previously chosen CQ Templates.	126
5.17	Stage 2: CQ Template Instantiation.	127
5.18	Stage 3: Reviewal of instantiated questions and answers.	127
5.19	Questionnaire construct averages and standard deviations across all 6 par- ticipants where for System Trust and TMv1 Suitability, the higher score the better, and for Workload, the reverse is true; the lower the better.	133
5.20	Questionnaire construct averages and standard deviations without P3 where for System Trust and TMv1 Suitability, the higher score the better, and for Workload, the reverse is true; the lower the better.	133
B.1	Nodes and Edges of the Provided Schema Relations.	248
B.2	Nodes and Edges instantiated by Oscar Wilde.	249

Chapter 1

Introduction

1.0.1 Motivation

A Knowledge Graph (KG), as defined by Hogan et al. [1], is a structured graph that models real-world knowledge by representing entities as nodes and the relations between them as edges. A *node-edge-node*, otherwise known as *subject-property-object*, is called a triple, an example of which would be *Claire - lives in - Dublin*. A way to store this information on computers is through the use of the Resource Description Framework (RDF) [2]. This would translate the example triple above into the following (shown in the compressed format called TURTLE, TTL [3])

```
<http://example-ontology.com/Claire>  
  <http://example-ontology.com/livesIn>  
    <http://example-ontology.com/Dublin> .
```

Linking data in this way enables the construction of a web of knowledge that can transcend traditional information silos. An example of this can be found in the bridging of seven European data registries as carried out by the FAIRVASC project [4]. This enables clinical researchers to review aggregated patient information across countries giving insight into rare kidney disease. Another example is DBpedia [5], a large-scale KG (as of the 2016 release, containing 9.5 billion triples [6]) that contains encyclopedic information extracted from Wikipedia [7] information boxes. In spite of the clear benefits of representing information in KG format, unless the relevant people e.g. clinicians, literary scholars, or any lay user [8] (an ordinary web-user, i.e. an individual who does not possess technical knowledge of RDF-related technologies) can explore them, such benefits may not readily surface.

The KG use case that this thesis will be focusing on is one that sits in the digital humanities domain, a domain that can be defined as the intersection between the *study of humanity* (which includes the disciplines of philosophy, English language and literature, religion, and history) and computers [9]. The KG in question is the Virtual Record Treasury of Ireland (VRTI) KG [10] which is a historical KG that has been created by historians and computer scientists as part of the VRTI project [11]. The VRTI project aims to digitally reconstruct records that were lost in the destruction of the Public Record Office of Ireland during the Irish Civil War in 1922. The VRTI KG stores and links people, places, and

more recently, events (entities) that are identified in those records; it also links entities and records, and utilises a variety of different ontologies in order to do so. Ontologies provide the specific structured language required for information about a KG’s schema to be stored in a way that is interpretable by both humans and computers. The VRTI KG makes use of, and at times extends, CIDOC CRM [12], a widely adopted cultural heritage ontology standard, GeoSPARQL [13], a standard for representing and querying geospatial data, and its own custom VRTI ontology [14]. The resulting KG is richly interconnected and semantically expressive, but its richness also brings complexity.

While the potential of KGs is enormous, there exist significant barriers to entry between the KG and lay users. KGs tend to be complex, large, and require a level of technical expertise in order to query over them [15]. Therefore, for the purpose of this thesis, a complex KG is defined as one whose underlying ontology introduces semantic and structural complexity that makes the graph difficult to interpret and explore without prior ontological or technical expertise. These barriers are exemplified by the VRTI KG: it is made up by several ontologies and vocabularies, and contains over 2.9 million triples. Like many RDF-based KGs, directly querying it requires knowledge of SPARQL [16] and an understanding of the KG’s schema. Lay users therefore face a clear orientation barrier when first engaging with a KG: they may not know what questions are possible, how to map those questions to the structure outlined in the ontology, or how to express them in a formal query language. This first-contact problem can prevent users from taking their initial steps in exploration and, unless mitigated, risks leaving large-scale humanities KGs inaccessible to the broader communities they aim to serve.

This thesis defines the first-contact, *"where to begin"* question as the ***Initial Exploration Problem (IEP)*** and characterises it through a conceptual framing comprising of three interrelated barriers:

1. **Scope Uncertainty barrier:** No prior awareness of what questions the KG can answer. This is not simply a matter of vague search intent or query refinement, as in traditional exploratory search [17], nor is it addressed by standard usability solutions. Rather, it is a problem of orientation: the user may not know what kinds of questions are answerable or where to begin their exploration.
2. **Ontology Opacity barrier:** Inability to interpret class/property terms and hierarchies (e.g., abstract event types, such as `E13_Attribute_Assignment` or `E65_Creation` from the CIDOC CRM ontology [12]), which greatly hinders exploration [8, 18]
3. **Query Incapacity barrier:** Lack of skills to formulate structured queries using formal query languages such as SPARQL [16, 19].

More concretely, it is possible to understand the problem as the following:

The Initial Exploration Problem (IEP) is the temporally bounded, pre-goal state experienced by a lay user at first contact with an unfamiliar, complex KG, characterised by the simultaneous presence of scope uncertainty, ontology opacity, and query incapacity, which together prevent the user from identifying a meaningful starting point for exploration.

In the following subsections, KG exploration solutions will be examined through the lens of addressing the barriers characterised in the IEP framing. These barriers are inherently overlapping, and individual solutions often address more than one. However, the core limitation across existing approaches is not the absence of partial solutions, but the absence of approaches that addresses all barriers simultaneously.

1.0.1.1 Scope Uncertainty and Ontology Opacity

A large number of interfaces have been developed to support exploration over KGs, but these typically fail to address the fundamental issues of scope uncertainty and ontology opacity that are characterised in the IEP framing. Faceted browsers and search interfaces, such as SampoUI [20], require users to already know enough about the domain to make the first meaningful decision. They do not present a clear or pedagogically informed starting point, and users are instead expected to choose a facet or construct a search term without any understanding of what the KG can or cannot answer. In these systems, the actual scope of the KG is only revealed after a user has already begun interacting with it, which presupposes the very knowledge the user is likely to be missing. As a result, faceted search places the cognitive burden of orientation on the user rather than providing support at the moment it is most needed.

Graph summary and overview visualisation tools, such as ViziQuer [21], attempt to give users a high-level view of the structure of a graph. While these tools can be powerful for technically experienced users, they are often overwhelming for non-technical audiences, particularly in the case of large or semantically dense KGs such as those built in the humanities. Visualisations that attempt to summarise millions of triples inevitably introduce visual and cognitive complexity, and users can struggle to extract meaningful starting points from these views. Moreover, the diagrams themselves are often semantically opaque: they display node and edge types defined in complex ontologies, leaving non-expert users unable to interpret what the structures mean or how they relate to the questions they wish to ask. Thus, while visual summaries aim to provide orientation, they frequently reproduce the problem of ontology opacity rather than alleviating it.

1.0.1.2 Query Incapacity

Several systems seek to lower the barrier of query formulation, but these solutions focus primarily on the mechanics of translating natural language into SPARQL rather than helping users understand what questions are possible to ask in the first place. Natural language to SPARQL interfaces, such as KG-GPT [22] or large language model (LLM) - based pipelines for generating SPARQL over federated KGs [23], best illustrate this. While they remove the burden of writing SPARQL syntax, they still rely entirely on the user knowing what to ask. If a user has no initial sense of the KG’s scope (what classes, properties, or event types exist) they cannot form the natural-language questions (NLQs) that these systems expect. In this way, such interfaces address query incapacity but leave scope uncertainty, as characterised in the IEP framing, untouched. As a result, they become useful only after the user has already overcome the hardest part of the exploration

process.

Visual query builders, such as KGViz [24], offer users a drag-and-drop interface to construct queries, but they too depend on prior understanding of the underlying schema. A user must know which classes or relationships are appropriate for the query they wish to form, and visual composition does not remove the semantic burden associated with choosing the correct building blocks. This again illustrates ontology opacity: even when the interface hides SPARQL syntax, it cannot hide the conceptual structure of the ontology itself. Consequently, these tools reduce some technical friction but still require users to possess a level of schema awareness that lay users are unlikely to have when first approaching a KG.

1.0.1.3 Starting Point Solutions

Yet other work attempts to provide users with a starting point, but these solutions tend to be either too technical, too reactive, or too narrow in scope to meaningfully address the IEP. Pre-written SPARQL queries, for instance, are often included as part of KG documentation or technical demonstrations [25]. While these queries illustrate what the KG is capable of returning, they are rarely designed with pedagogical intent. They assume that (1) the user is capable of understanding SPARQL and (2) that they can mentally map a technical query onto the domain concepts they care about. They provide little guidance in understanding why a query is meaningful, how it relates to the KG’s structure, or what broader set of questions might be possible.

Recommendation systems [26] represent another class of "starting point" solutions, but they suffer from a fundamental mismatch with the IEP. Much like the cold-start problem [27] in recommender systems, a user with no previous engagement provides no basis upon which recommendations can be meaningfully generated. In the KG context, the user is "cold" not because their behavioural history is missing, but because their cognitive orientation is missing: they do not yet know what the KG contains, nor how to ask for information within it. Scope uncertainty is therefore analogous to the user’s cognitive cold start. Any recommendation mechanism that relies on user activity or relevance feedback is inherently reactive and can only become useful once the user has already begun exploring. It cannot address the core challenge of helping users take that first step.

OnSET [28] provides a different approach offering a topic-guided visual query building interface. Although it does not explicitly address the IEP, it implicitly attempts to tackle aspects of the problem as characterised in the IEP framing by reducing ontology opacity: by surfacing topics and visual structures, it helps users build a mental model of the KG’s underlying schema. However, OnSET does not directly address scope uncertainty, as characterised in the IEP framing, as users still need to choose which topic to begin with. It also does not fully mitigate query incapacity, as users ultimately must understand how to construct a query using the visual interface. In this sense, OnSET offers a structural entry point into a KG that supports open-ended exploration, leaving users to decide how to navigate the space.

Pre-formed Natural Language Question and Answer entry-points (NLQAs) provide

an intuitive and cognitively aligned way for lay users to begin exploring a KG. Unlike faceted browsing, visual summaries, or query-building interfaces, which often presuppose some familiarity with the KG’s structure, questions allow users to start from familiar linguistic expressions. A well-formed question reveals the scope of the KG explicitly by signalling, in plain language, that a particular kind of information *can* be asked about and the answer can bring the user to a new area of the KG without the requirement of knowing SPARQL [16]. For example, a question such as "Where was Oscar Wilde born?" implicitly communicates that the KG contains people, birth events, and associated places, even if the user has no prior understanding of the underlying ontological structures or class hierarchies that store this information. NLQAs also introduce ontology concepts implicitly. Terms such as "birthplace", "associated place", or "related event" map onto complex ontology patterns, such as CIDOC CRM’s event modelling conventions, without requiring the user to understand these patterns directly. In this way, questions and answers (e.g., question: *"Where was Oscar Wilde born?"* Answer: *"Dublin, Ireland"*) serve both as entry points and as subtle scaffolding tools: they bridge a user’s natural mental model of the domain with the formal semantic structures encoded in the KG. Furthermore, the answers of NLQAs bring users to areas of the KG that they can then explore out from, i.e., fulfilling the role of an "entry-point" into a KG.

Crucially, NLQA entry-points do not require any knowledge of query languages. A user does not need to know SPARQL, nor do they need to understand how subject-property-object triples form the building blocks of a KG. This makes questions a particularly effective means of addressing query incapacity, as they shift the burden of formalisation away from the user and onto the system or interface. **However, such NLQA pairs must be carefully curated so that as a set they present the potential of a KG (scope), and individually they are grounded in the KG’s current contents, are clearly worded, and are likely to bring users to areas of the KG of interest.** In other words, such NLQAs are not just expected to function as usability features but also as epistemic communicators. This is where domain expertise becomes essential as historians or other domain experts are the ones who understand which NLQAs would fit the criteria of an entry point into a specific KG. This solution, therefore, introduces a secondary human-computer interaction (HCI) challenge: how can we support the domain experts tasked with creating such NLQAs?

In addressing the secondary HCI challenge, two further considerations are surfaced: scale and technicality. Humanities KGs often contain thousands of people, places, and events, making it infeasible for domain experts to manually inspect or recall the KGs contents in order to author a comprehensive set of NLQAs. Along with this, many of the challenges that are present for lay users appear here also: while domain experts may not necessarily encounter the problem of where to start, they likely lack the technical expertise to query over KGs. Taken together, these considerations point toward the need for a multistage framework that supports domain experts in producing curated NLQA entry points at scale. The research presented in this thesis assesses the extent to which Natural Language Question and Answer entry-points (NLQAs) address the Initial Explo-

ration Problem (IEP), as characterised in its framing, and introduces and evaluates such a multistage NLQA curation framework which operationalises these ideas with the aim of mitigating the IEP in the humanities domain using the VRTI KG as a use case.

1.0.2 Research Questions

The research presented in this thesis is driven by the following research questions (RQs):

RQ1:

"To what extent can domain expert curated natural language question-answers (NLQAs) as entry points to RDF-based Knowledge Graphs (KGs), be perceived to mitigate the Initial Exploration Problem (IEP), as characterised in its framing, for lay users in the humanities domain?"

RQ2:

"To what extent does a multistage framework, when instantiated, support domain experts in the curation of such natural language question-answers (NLQAs) as entry points to RDF-based Knowledge Graphs (KGs) in the humanities domain?"

Terms used within the RQs:

- **Domain Expert:** For this research, an individual with deep, scholarly knowledge of the content within a humanities KG (e.g., a historian) but who lacks the technical expertise to formulate formal queries (e.g., in SPARQL) against the KG. Their role is to perform the curation aspect of the framework.
- **NLQA:** Human-readable question and answer pairs that act as starting points for KG exploration. For this research, NLQAs are grounded in structured information retrieved from the KG (e.g., via query templates such as SPARQL), rather than being direct natural-language-to-query translations.
- **Mitigate the Initial Exploration Problem (IEP):** To reduce the perceived impact of the barriers characterised in the IEP framing – scope uncertainty, ontology opacity, and query incapacity – thereby improving the perceived understanding of the contents and approachability of the KG, enabling identification of an exploration starting point.
- **Curation:** The human-driven process of reviewing, refining, approving, or discarding candidate exploration starting point NLQAs to ensure clarity, relevance, and answerability.
- **Framework:** A conceptual, multistage structure that defines the stages and relationships between the stages required to support domain-expert curation of NLQA pairs, while remaining agnostic to the specific techniques or technologies used to instantiate those stages.

- **Instantiation:** The concrete operational realisation of a conceptual framework, in which abstract stages are realised using specific techniques and tools, enabling the framework to be executed, observed, and evaluated in practice. An instantiation adheres to the structural commitments of the framework while remaining one of many possible realisations.

1.0.3 Research Objectives

These research questions can be decomposed into the following research objectives (ROs):

- **RO1:** To identify and characterise the barriers lay users face when beginning to explore unfamiliar complex KGs in the humanities domain (the barriers collectively characterise the Initial Exploration Problem (IEP)).
- **RO2:** To identify and specify the lifecycle of NLQA entry points into KGs, and to design, instantiate, and implement a multistage framework over a complex digital humanities KG to support domain expert curation.
- **RO3a:** Lay User Impact: To measure the extent to which curated NLQA entry points are perceived to mitigate the IEP, as per its framing, for lay users in the humanities domain.
- **RO3b:** To technically investigate and analyse Stage 1 and Stage 2 of the framework through its instantiation.
- **RO3c:** Domain Expert Evaluation: Evaluate with domain experts the viability of the instantiated framework.

1.0.4 Research Methodology

1.0.4.1 Problem Identification and Conceptualisation

The first step of this research involved a review of existing KG exploration interfaces. This review revealed a common gap: existing tools primarily assist users who already have a query in mind. In other words, they have not been designed to orient a new user. This gap surfaced the problem conceptualised in this thesis as the IEP: the problem of supporting a lay user to begin exploration of a KG. To the best of the author’s knowledge, prior to this research, there was no unifying term or characterisation of the problem of first contact with an unfamiliar KG in literature. Instead, aspects of the problem appeared in fragments across prior work: the cognitive demands of sensemaking [29], the practical difficulties of SPARQL [30], and the opacity of ontology structures [31].

By naming the IEP and developing its conceptual framing, these interrelated barriers are unified under a single analytical construct thereby addressing RO1. Although individual aspects of the IEP have been discussed previously, such as the difficulty of composing SPARQL queries [19] and the cognitive load of interpreting ontology hierarchies [32], these works typically addressed fragments of the problem rather than its defining moment: the user’s first encounter with the graph. By naming the IEP and developing

its characterisation, this research provides both a vocabulary for discussing the orientation barrier explicitly, and a framework for designing and evaluating targeted interventions at the precise moment when a lay user must decide how to begin exploration.

While the IEP is domain-agnostic, this thesis will focus on its impact in the humanities domain, a domain where the problem of *where to start* can be felt acutely due to the often complex nature of the KGs in question [10, 33].

1.0.4.2 Technical Approaches

Tús Maith (/tu:s "mah/), Irish for "a Good Start", addresses RO2 through a structured human-in-the-loop framework. This framework decomposes the problem of creating NLQA starting points into three stages: (1) **CQ Template Derivation**: deriving Competency Question (CQ) templates [34] from the KG (CQ templates), (2) **CQ Template Instantiation**: instantiating these templates with specific, relevant entities from the KG (filled-in questions and answers), and (3) **Expert Curation**: enabling expert curation to produce final, user-facing NLQAs called Curated Questions and Answers (CuQAs). CQs [34] are questions typically designed to guide ontology development or evaluate knowledge completeness [35]. However, in this work, they are adopted as a mechanism for supporting user exploration. The structured approach introduces the following NLQA lifecycle, as numbered and shown in Figure 1.1:

1. **Competency Question (CQ) Templates**: A general template question derived from the structure of the KG, e.g. *"Where was a person born?"*.
2. **Filled-in Question and Answer**: A CQ template filled in with popular entities within the graph and/or entities the domain expert feels may be of interest to produce a question-answer pair, e.g. *"Where was Oscar Wilde born? Dublin"*.
3. **Curated Question and Answer (CuQA)**: The filled-in question and answer accepted or refined by a domain expert for presentation to users to act as starting points into the KG.

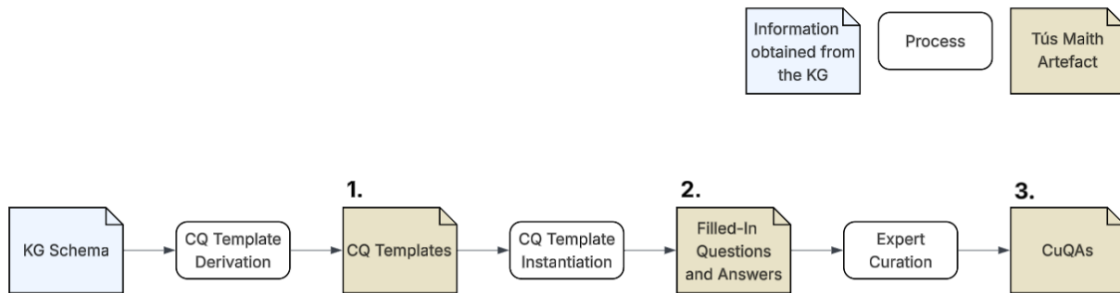


Figure 1.1: The Tús Maith framework: A framework to create CuQAs.

The decision to derive CQs from the schema information extracted from instance-level data rather than from the full ontology was intentional. Ontologies can provide more terms

than are currently being used by the KG and this work aims to support exploratory access to available content, rather than exhaustively covering the ontology itself.

Rather than translating CuQAs into bespoke SPARQL queries, the framework anchors each CuQA to specific KG entities. User interaction triggers the materialisation of the entity’s semantic neighbourhood, within which the question’s answer is presented as a particular perspective over the broader KG context. This shifts the role of questions from mechanisms of answer retrieval to guided entry points into the structure and scope of the KG, directly addressing the challenge of how users begin exploration in the absence of well-formed information needs. While KGs and those interested in exploring them may be multilingual, the work conducted as part of this thesis focused on English-language CuQAs.

In order to evaluate the Tús Maith framework it was instantiated as Tús Maith version 1 (TMv1) through the utilisation of Large Language Models (LLMs) in Stage 1, CQ Template Derivation, and Stage 2, CQ Template Instantiation. TMv1 can be seen in Figure 1.2 and a UML activity diagram describing the process by which CuQAs are generated can be seen in Figure 1.3.

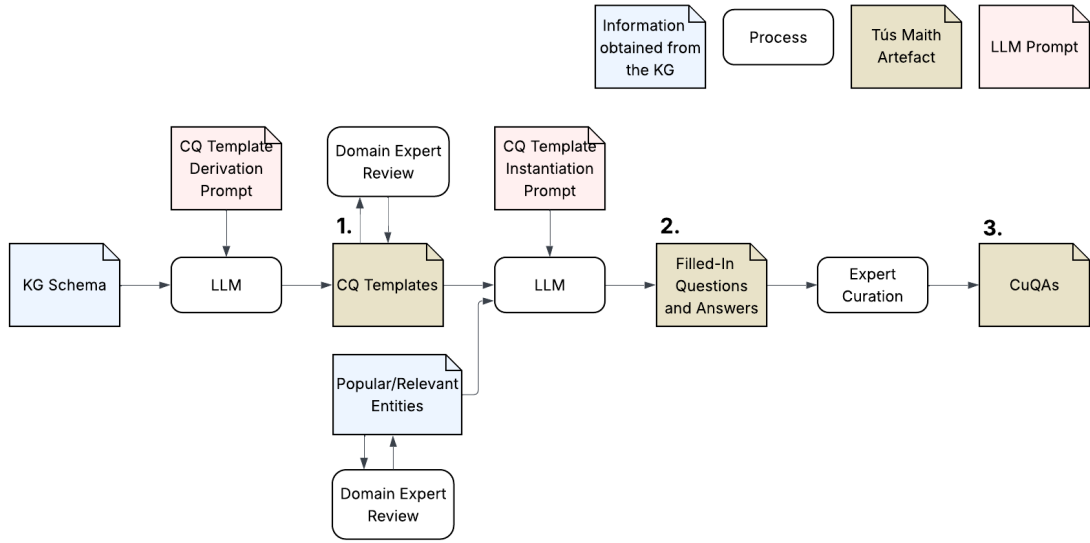


Figure 1.2: TMv1: An LLM-supported instantiation of the Tús Maith framework.

LLMs were chosen as the means of instantiating the framework for the reasons outlined in Table 1.1.

While LLMs are a scalable, flexible, and generalisable method of instantiating stage 1 and stage 2 of Tús Maith, they bring their own set of challenges and considerations. Their inherent "black-box" nature raises concerns regarding explainability and trustworthiness [36]; it is difficult to audit their reasoning for generating a specific question template or entity pairing. Furthermore, operational factors such as financial cost, computational latency, and environmental impact present practical constraints on the sustainability of the instantiation. Therefore, it was imperative that the system-level design of TMv1 did not treat the LLM as an infallible oracle, but rather as a powerful yet constrained component,

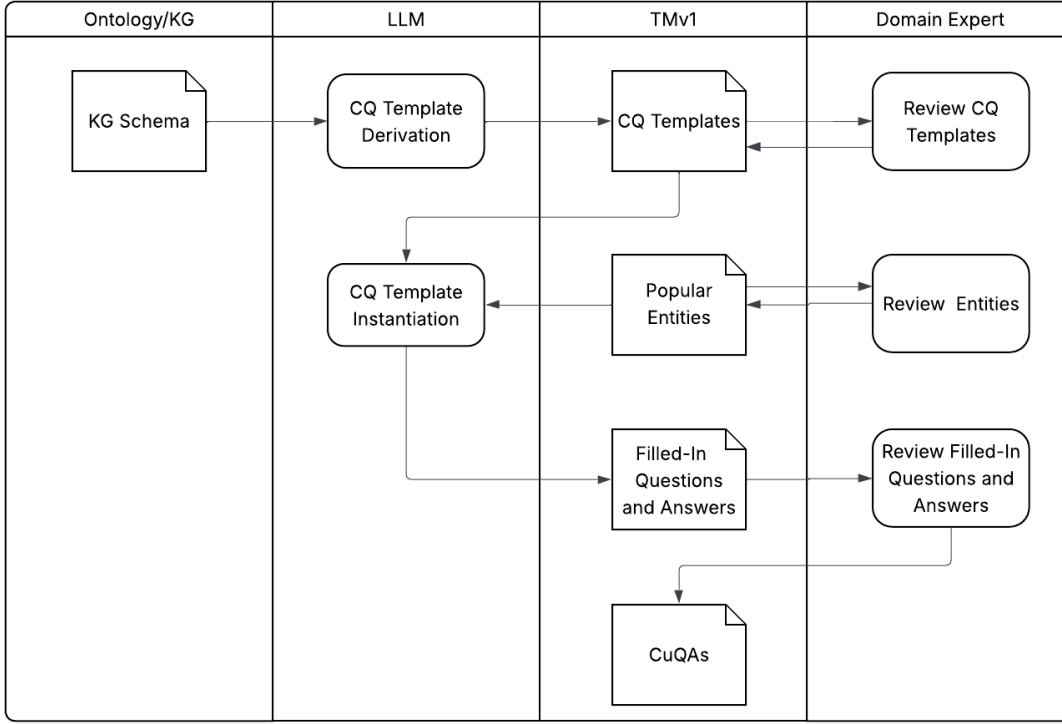


Figure 1.3: TMv1: A UML activity diagram denoting the process by which CuQAs are generated.

with these considerations shaping the human-in-the-loop curation system.

In order to evaluate with lay users the use of CuQAs as an entry point to an unfamiliar KG in-situ, a new feature was integrated into the VRTI Explorer [37], an interface that has been already deployed to sit between the VRTI KG and the end user enabling the user to explore the VRTI KG¹.

All code and datasets associated with this thesis can be found in the following GitHub repository: https://github.com/mcnamacl/T-s_Maith_Thesis_Repository.

¹<https://kg.virtualtreasury.ie>

Table 1.1: Comparison of options for instantiation of stage 1 and stage 2 of the Tús Maith Framework.

Dimension	Human	Programmatic	LLMs
Scalability	Manual, limited to domain experts	Template-dependent	Highly scalable, rapid generation
Coverage	Expert-bounded	Rule-constrained	Flexible, generalisable
Language Quality	Domain-appropriate	Often rigid, can be unnatural	Natural, fluent
Accuracy Control	Inherent	Deterministic	Requires verification

1.0.5 Evaluations

In order to support the investigation of RO3a - RO3c, five evaluations and one proof-of-concept were carried out. Those five evaluations comprised of two LLM-supported experiments and three user evaluations. RO3a involved two of the user evaluations. The first of these was an exploratory study to investigate whether the use of CuQAs as a method of entry into a KG was a promising approach (User Evaluation 1), and the second larger study looked specifically at the extent to which CuQAs address the barriers characterised in the Initial Exploration Problem (IEP) framing (User Evaluation 2). RO3b required technical evaluation of the extent to which LLMs could successfully function as CQ template derivators (LLM Experiment 1) and in CQ template instantiators (LLM Experiment 2). The final fifth evaluation addressed RO3c and was a domain expert user evaluation that looked to assess the extent to which the Tús Maith framework is useful to domain experts in creating publishable CuQAs (User Evaluation 3). All five evaluations were undertaken using the VRTI KG. The proof-of-concept was carried out in support of Major Contribution 1 presented below with LetterSampo [38]. This proof-of-concept demonstrates the Tús Maith framework’s generalisability beyond the VRTI KG with a similarly sized but different humanities KG.

The three user-centred evaluations employed mixed methods analysis. RO3a consisted of user evaluation 1 and 2, and RO3c, user evaluation 3. All three evaluations are discussed in Chapter 5.

1.0.5.1 User Evaluation 1

This was the first user-focused experiment and it was exploratory in nature. It aimed to understand whether CuQAs showed promise as entry points into a KG. At this stage in the research, Tús Maith consisted of 2 features - CuQAs and commonly searched key words, that worked in conjunction with a search bar. CuQAs and commonly searched key words were integrated into an early version of the VRTI Explorer already containing the search bar and users were asked to evaluate their experience interacting with the VRTI Explorer with the 2 added features. The study involved 12 participants with varying levels of background knowledge of the VRTI project who carried out the evaluation asynchronously via an online questionnaire. The results of the evaluation were analysed using quantitative and qualitative measures and indicated that most users with limited prior knowledge of the VRTI Project (3 out of 4) preferred CuQAs as their method of beginning exploration of the KG. These results prompted further investigation into the efficacy of CuQAs.

1.0.5.2 User Evaluation 2

User evaluation 2 formed the second iteration of the assessment of CuQAs in-situ. By this evaluation, Tús Maith had been streamlined from key words and CuQAs to just CuQAs, based upon the results of User Evaluation 1 and the IEP characteristics had been defined as scope uncertainty, ontology opacity, and query incapacity. Based on this, a

questionnaire was designed drawing inspiration from the following existing questionnaire paradigms: PSSUQ [39], SUS [40], and TAM [41], and also containing custom questions. A set of CuQAs were obtained from two expert historians working on the VRTI Project - three CuQAs for person and three for place. Those CuQAs were implemented as a feature extending the latest version of VRTI Explorer. The objective of the evaluation was to assess the extent to which CuQAs were perceived to mitigate the problem named as the IEP; therefore, the questionnaire contained questions to be answered pre-interaction with the VRTI Explorer with the CuQAs feature and questions to be answered post-interaction. The questions were designed such that they mapped to two of the three observable IEP characteristics (scope uncertainty and ontology opacity). By the nature of the CuQA implementation using the VRTI Explorer, query incapacity was not directly observable because the interface deliberately abstracted away the underlying query mechanisms. Participants were therefore unable to experience or express query incapacity directly. Instead, the evaluation focused on users' perceived ability to orient themselves and understand the scope and structure of the KG. The evaluation involved the recruitment of participants with limited familiarity of the VRTI Project and VRTI Explorer but who were identified to have an interest in history. The resulting participants therefore consisted of 35 students studying history in Trinity College Dublin across multiple different class cohorts. Like user evaluation 1, the participants carried out the evaluation tasks and filled in an associated questionnaire online asynchronously. As the IEP is conceptualised as an orientation problem compounded by technical problems (SPARQL, complex ontologies [12, 13]), perceived usefulness, perceived clarity, and perceived confidence were considered appropriate measures mitigation. The results of the evaluation indicated that CuQAs statistically are perceived to mitigate two of the three barriers characterised in the IEP framing (scope uncertainty and ontology opacity). While not being shown to increase orientation pre- and post-use of CuQAs, there are plausible considerations for why that was the case (high baseline confidence due to domain familiarity).

1.0.5.3 User Evaluation 3

The final user evaluation was a full end-to-end evaluation of TMv1 by six VRTI domain experts. In order to conduct such an evaluation, TMv1 was implemented as a prototype interface enabling the historians to use the three stages of the Tús Maith framework in practice. The evaluation was conducted online and made use of a combination of behavioural, quantitative, and qualitative metrics. Two design implications arising from LLM Experiment 2 were incorporated and explored within this evaluation namely: the provision of provenance information for the filled-in NLQAs and LLM model selection transparency (presentation of which underlying model was used). Qualitative feedback indicated that providing this provenance and model selection information calibrated user trust in TMv1. Overall, the results suggest that TMv1 is viable as a method of creating CuQAs, with all six participants creating NLQAs that they deemed suitable as entry points into the VRTI KG (CuQAs).

The next two experiments were LLM-supported and both experiments are discussed in

1.0.5.4 LLM Experiment 1

The objective of LLM Experiment 1 was to assess the extent to which an LLM could produce CQ templates of sufficient quality. The experiment employed both quantitative and qualitative evaluation methods using both automatically generated tests and domain expert evaluation. It involved asking a historian with expert domain knowledge of the VRTI KG to create a set of CQs that they felt best reflected the contents of the VRTI KG. Then two different LLMs were each provided with different versions of the classes and properties of the VRTI KG (the full list of the classes and properties, a manually reduced list keeping only the most semantically expressive and user-relevant terms while excluding overly technical entries, or the manually reduced list plus some relationship information between classes and properties). Four different prompting strategies were used and the generated questions sets produced by the LLMs were semantically compared using SentenceBERT [42]. Finally, the highest ranked semantically similar questions set produced by each LLM were then given to two different historians in order to assess the questions using three metrics that had been identified to be criteria for a CuQA: relevance, clarity, and answerability. The results of this evaluation indicated for the first time what would become a guiding principle for the rest of the research - that LLMs are scaffolds for human CuQA generation, not substitutes; that while they can produce some relevant, clear, and answerable CQ templates, they still require expert oversight. A detailed discussion of this evaluation can be found in Chapter 5.

1.0.5.5 LLM Experiment 2

LLM Experiment 2 investigated how effectively LLMs can perform batch instantiation of CQ templates to create filled-in questions and answers. Using the domain expert authored CQ templates created in LLM Experiment 1, four different LLMs were provided with the set of CQ templates and 10 entities (with information about them) at a time, and the LLM was asked to fill in the relevant CQs with the information provided. This evaluation involved the creation of a reproducible, multi-level evaluation pipeline that: (a) operationalised strict slot-level, semantic (SentenceBERT (SBERT) [42]), and token-level (BERTScore [43]) fidelity checks with rescue rules and calibrated thresholds; (b) implemented a realistic entity-ranking and batching strategy to produce curator-relevant silver-standard test sets; (c) compared two provenance conditions (requiring full trace-back of all information that was used to instantiate the CQ template vs. requiring just the URI of the entity whose information was used to instantiate the CQ template) across repeated zero-shot runs to measure impact on completeness, accuracy, hallucination, and efficiency; and (d) evaluated a practical JSON-validation proxy for operational contexts. Across approximately 25,000 generated filled-in questions and answers, the evaluation revealed that LLMs can effectively scaffold curator systems, though with important practical considerations. Model selection emerged as a critical trade-off of accuracy versus speed and cost. The JSON validator proved effective for factual questions but insufficient for reasoning-

based ones, confirming its role as a first-pass filter rather than a complete quality measure. Notably, full provenance logging significantly increased runtime without substantial accuracy gains, suggesting that requiring only basic provenance is more practical for batch workflows. These findings directly informed the system design of the TMv1 prototype, providing both empirical characterisation of LLM behaviours and concrete design patterns for the curator user interface (UI), while supporting the design principle identified in LLM Experiment 1, that LLMs are helpful assistants, not replacements for humans. The full experiment is described in Chapter 5.

1.0.5.6 LetterSampo Proof-of-Concept

The final assessment of Tús Maith, as instantiated through TMv1, was conducted using the LetterSampo correspondence KG [38]. This proof-of-concept is in additional support of Major Contribution 1 by examining whether the Tús Maith framework can be applied beyond the Virtual Record Treasury of Ireland (VRTI) KG to a separate, similarly sized humanities KG with a different ontology and data structure. The purpose of this proof-of-concept was not to provide a full evaluation of effectiveness comparable to the VRTI-based studies, but rather to assess the portability of the Tús Maith framework when instantiated in a new KG context. As TMv1 constitutes a concrete instantiation of Tús Maith, demonstrating that TMv1 can be applied to a structurally distinct KG without modification to the underlying framework provides evidence that the framework itself is not dependent on any single ontology, dataset, or institutional context. LetterSampo was selected as it represents a mature, large-scale humanities KG of approximately 2.9 million triples, comparable in scale but different in structure and domain emphasis to the VRTI KG. Applying TMv1 to LetterSampo therefore provides a meaningful test of whether stage 1 and stage 2 of the Tús Maith framework (the absence of access to domain expert precluded investigation of stage 3, expert curation) can generalise across heterogeneous humanities KGs. This proof-of-concept demonstrates that stage 1 and stage 2 of the Tús Maith framework can be successfully instantiated and executed over a structurally distinct humanities KG using the same conceptual stages and workflow as employed for the VRTI KG. In doing so, it provides supporting evidence for Major Contribution 1. The full proof-of-concept is presented in Chapter 5.

1.0.6 Thesis Contributions

This thesis presents one major and two minor contributions. The major contribution is **Tús Maith**: the concept of CuQAs and a multi-step approach for producing, curating, and publishing them as entry points into a KG. The first minor contribution is the identification, definition, and characterisation of the IEP, and the second, is the empirical evaluation of LLM assisted scaffolding.

Publications The following peer reviewed publications are associated with the contributions and are presented below:

McNamara, C., Hederman, L., & O’Sullivan, D. (2025). **Mind the Gap: LLMs, Competency Questions, and the Non-Technical User in the Humanities Domain**. In First International Workshop on Users and Knowledge Graphs @ SEMANTiCS 2025 [44].

- This paper introduces and characterises the IEP, further discussion of which is presented in Chapter 3: The Initial Exploration Problem (IEP) and Related Work.
- Along with introducing the IEP, this paper also introduces Tús Maith as a three stage framework (see Chapter 4) and the experiment described in Chapter 6.

McNamara, C., Hederman, L., & O’Sullivan, D. (2026). **Filling the Gap: LLMs as Scaffolds for Competency Question Instantiation**. In Proceedings of the 31st International ACM Conference on Intelligent User Interfaces (IUI ’26) Association for Computing Machinery, New York, NY, USA, 822–835. <https://doi.org/10.1145/3742413.3789141> (Acceptance rate: 20.3%, 114/561 submissions) [45].

- This paper details LLM Experiment 2 that can be found in Chapter 5, Section 5.8.
- It further translates the empirical findings of the experiment into system design decisions for TMv1.

Submission Under Review The following journal article has been submitted to the Journal on Computing and Cultural Heritage, JOCCH², and is under peer review:

McNamara, C., Hederman, L., & O’Sullivan, D. (2026). **Curated Questions and Answers as Exploratory Entry Points into Cultural Heritage Knowledge Graphs**. Journal on Computing and Cultural Heritage (JOCCH).

- This article details presents both User Evaluation 2 (detailed in Chapter 5, Section 5.6) and User Evaluation 3 (detailed in Chapter 5, Section 5.9).

Preprint The following preprint is live on arXiv:

McNamara, C., Hederman, L., & O’Sullivan, D. (2026). **The Initial Exploration Problem in Knowledge Graph Exploration**. arXiv preprint arXiv:2602.21066. [46]

- This preprint centres on the conceptualisation of Initial Exploration Problem (IEP) and where it sits within related literature. The content in this paper reflects the content of Chapter 3: The Initial Exploration Problem (IEP) and Related Work.

1.0.6.1 Major Contribution: Tús Maith

The major contribution of this thesis is the design, instantiation, implementation, and evaluation of the Tús Maith framework. Tús Maith operationalises the idea that curated NLQAs (CuQAs) can act as intuitive cognitive and linguistic scaffolds for lay users, revealing both the scope and structure of a KG without requiring prior knowledge of its underlying ontology or query language.

²<https://dl.acm.org/journal/jocch>

The framework comprises of three stages (as were presented in Figure 1.1): (1) the generation of competency question (CQ) templates from a KG; (2) the instantiation of these templates with popular or relevant entities extracted from a KG; and (3) domain expert curation, through which validated questions are reviewed, refined, and published as user-facing entry points for exploration. While CQs are traditionally employed as design-time artefacts for ontology engineering, Tús Maith repurposes them as exploratory instruments, recognising and operationalising their implicit ability to expose the scope and structure of a KG for lay users.

Through iterative empirical evaluation, this research provides evidence of the extent to which CuQAs are perceived to mitigate the Initial Exploration Problem (IEP) as per its framing by reducing scope uncertainty, demystifying ontology structure, and removing the need for SPARQL query formulation. This research further provides evidence that the three-stage framework (Tús Maith) to generate CuQAs that are grounded in a KG’s contents is considered viable by domain experts, the intended users of Tús Maith. In doing so, Tús Maith constitutes both a conceptual and practical advancement: conceptually, it demonstrates that early-stage KG exploration can be supported through curated linguistic entry points without requiring prior ontology comprehension or query formulation; practically, it delivers an implementable and transferable framework for creating such entry points across domain-specific KGs.

1.0.6.2 Minor Contribution 1: The Identification and Characterisation of the Initial Exploration Problem (IEP)

A theoretical contribution of this thesis is the identification, definition, and characterisation of the problem named the Initial Exploration Problem (IEP): the challenge faced by lay users in attempting to begin exploration of an unfamiliar KG. The problem named as the IEP is characterised in its framing as comprising of three interrelated barriers: scope uncertainty, ontology opacity, and query incapacity. By unifying these challenges under one heading, this thesis provides a shared conceptual vocabulary for analysing the first-contact experience and for designing interventions that specifically target the moment of orientation.

1.0.6.3 Minor Contribution 2: Empirical Evaluations of LLM-Assisted Scaffolding for KGs

The second minor contribution stems from empirical studies investigating how LLMs can assist in the creation of scalable, semantically accurate question scaffolds for KGs.

2a. Automated and Scalable Generation of CQs: This research provides further evidence to existing literature [47, 48, 49, 50] that LLMs can be used to generate CQs from the classes and properties of a KG. However, what it also shows is that such automatically generated CQs can be used for exploration purposes as opposed to ontology engineering ones. In doing so, they can increase coverage, producing a breadth of potential question templates that would be infeasible for human experts to construct retrospectively at scale. This demonstrates how automated scaffolding can complement expert curation by

broadening the exploratory surface of a KG.

2b. Evaluation of Entity-Fidelity-Constrained Generation: A more focused empirical contribution assesses how well LLMs perform when filling question templates to produce filled-in questions and answers with structured KG entity data under strict constraints. These constraints include (i) zero tolerance for hallucination, (ii) entity fidelity (ensuring that the filled-in question accurately represents real KG content), and (iii) prompt compliance. Evaluating LLMs in this tightly controlled context provides valuable insight into their behaviour in structured, data-driven applications, a setting where factual accuracy, ontological alignment, and interpretability are vital. A further contribution lies in the implementation and evaluation of a first-pass, lightweight validator that, through the utilisation of URI provenance, obtains 88-99% precision (depending on the LLM chosen) when compared to a ground truth (GT)-dependent semantic validator.

1.0.7 Thesis Overview

1.0.7.1 Chapter 2: Background

Chapter 2 presents the background information necessary to understand the technology used as part of the research presented in this thesis.

1.0.7.2 Chapter 3: The Initial Exploration Problem (IEP) and Related Work

Chapter 3 characterises the IEP through its framing and synthesises literature relating to the challenges identified as part of the problem.

1.0.7.3 Chapter 4: The Design, Instantiation, and Implementation of the Tús Maith Approach

This chapter provides the detail on the design of the Tús Maith framework, the instantiation of Tús Maith producing TMv1, and the implementation of TMv1 using the VRTI KG. It adopts a design-justified approach presenting relevant literature at the point at which it motivates or constrains a specific decision relating to Tús Maith and TMv1.

1.0.7.4 Chapter 5: Evaluations

Chapter 5 presents all five of the evaluations outlined in Section 1.0.5. It also outlines the proof-of-concept carried out to demonstrate the Tús Maith framework’s portability to a structurally different humanities KG. Finally, it synthesises the results across all evaluations.

1.0.7.5 Chapter 6: Conclusion and Future Work

The final chapter presents the conclusion of this thesis and future work identified as part of the research.

Chapter 2

Background

This chapter presents background and introductory material on key technical areas, intended to aid the understanding of the thesis for those unfamiliar with those areas. Specifically, it outlines large language models (LLMs), which are the specific technology used to instantiate Tús Maith to become TMv1, the ontologies that underpin Knowledge Graphs (KGs) in the digital humanities domain, and existing user interfaces (UIs) for digital humanities KGs.

2.1 Large Language Models (LLMs)

Large language models (LLMs) are utilised in both the instantiation of Tús Maith forming TMv1 (decoder-only LLMs) and, as mechanisms to compute semantic similarity between generated text and a ground truth (GT) text for the two technical evaluations (LLM Experiment 1 and LLM Experiment 2) (encoder-only LLMs).

2.1.1 Architecture

What differentiates LLMs from previous neural network (NN) architectures like Long-Short-Term-Memory (LSTM) or Recurrent-Neural-Network (RNN) models is their use of the self-attention mechanism combined with positional encoding, and a fully parallelisable architecture [51]. Rather than processing each token sequentially, in self-attention, each token is projected into query, key, and value vectors. Attention weights are computed by comparing queries with keys across the sequence, and these weights are used to aggregate value vectors. This is what allows all input tokens of the LLM to attend to all other tokens in parallel building context and mitigating limitations associated with modelling long-range dependencies as encountered by sequential NNs like RNNs. LLMs still however are limited by the size of their context window, the number of tokens that the LLM can attend to at once. This constraint limits the extent to which large structured resources such as KGs can be provided to a model within a single prompt, necessitating selective retrieval, summarisation, or decomposition of graph content when LLMs are applied to KG-based tasks. A full transformer consists of both an encoder and decoder and as such full transformer-based models are primarily used for sequence-to-sequence tasks,

e.g., translation: the input is required to be encoded to understand what exactly is being asked to be translated, and then decoded, transforming or translating it into the target language. In contrast, generative LLMs, like the models used as part of this thesis (e.g., Open AI’s GPT-4o [52] and GPT-5 [53] or Google’s Gemini 2.0 Flash [54]) are decoder only, optimised for autoregressive text generation. The encoder-only models, like the ones used for part of this thesis (e.g., BERT-style architectures like RoBERTa-Large [55]), are optimised for text classification and representation learning which enable them to compute the semantic similarity between two pieces of text.

Decoder-only LLMs generate text autoregressively by predicting the next token conditioned on prior context and the input prompt. As a result, outputs are probabilistic and sensitive to prompt structure, ordering, and contextual framing. When used for structured generation tasks such as competency question (CQ) generation over complex KGs, these models do not inherently guarantee structural validity or schema compliance. Consequently, prompt design and external constraints play a critical role in guiding output consistency and answerability.

Encoder-based transformer models generate dense vector embeddings representing semantic content. In this thesis, generated text and GT text are embedded in a shared vector space and compared using cosine similarity, thus enabling large-scale comparison of semantic alignment between generated and GT text while avoiding reliance solely on surface-level metrics e.g., lexical overlap.

2.2 Ontologies

This section details the primary ontology used to model the information contained within Virtual Record Treasury of Ireland (VRTI) KG and the structure of VRTI KG itself.

2.2.1 CIDOC Conceptual Reference Model (CIDOC CRM)

The CIDOC Conceptual Reference Model (CIDOC CRM) [12] is an ISO-standard event-modelling ontology that is designed to support the integration, mediation, and interoperability of heterogeneous cultural heritage information. It is an extensive and complex ontology that enables historical information, like the kind curated as part of the VRTI project, to be comprehensively represented in RDF format.

Unlike entity-centric data models, CIDOC CRM uses an event-centric modelling paradigm in which relationships between entities are typically mediated through events. For example, rather than directly asserting relationships between people, such relationships are typically represented through events (e.g, a birth event linking a person, time-span, and place). This enables rich contextual representation but introduces structure complexity contributing to **ontology opacity**. CIDOC CRM further consists of a large set of abstract classes (e.g., `E21_Person`, `E53_Place`) and properties that can be combined to model complex historical relationships. Its high level of abstraction and extensibility allows heterogeneous archival and historical information to be integrated within a shared semantic structure. However, such abstraction once again makes it challenging for lay users to interpret the underlying

ontology (**ontology opacity**).

2.2.2 The Virtual Record Treasury of Ireland (VRTI) KG

The Virtual Record Treasury of Ireland (VRTI) KG [10, 56] is a large-scale cultural heritage KG developed as part of the VRTI project, an international digital humanities initiative established to reconstruct archival material lost in the fire that claimed the Public Record Office of Ireland in 1922 [57, 11]. The destruction of the Public Record Office resulted in the loss of a substantial portion of Ireland’s documentary heritage, including census returns, court records, wills, and administrative documentation. The VRTI project addresses this loss by digitally identifying, aggregating, and reconstructing surviving copies, transcripts, and references from those records from repositories across Ireland and beyond.

Within the VRTI KG, CIDOC CRM forms the core ontological framework used. Additional vocabularies include its own custom VRTI ontology [14], GeoSPARQL [13], and *logainm*¹. The use of multiple complex ontologies to model the VRTI KG further exacerbates **ontology opacity**. An example of an entry for a person entity within the KG is as follows: a historical person is modelled as `crm:E21_Person`. Their birth is not recorded as a direct attribute of the person but instead, birth is modelled as a distinct event (`E67_Birth`) that brings the person into existence. The birth event is then linked to a time-span and place, while the person is separately linked to names and identifiers. How this appears functionally in the KG using Oscar Wilde as an example can be seen in Listing 2.2.1.

Listing 2.2.1: Turtle (TTL) Representation of Oscar Wilde’s Birth Record in the VRTI KG using CIDOC CRM and VRTI ontologies.

```
1 @base <https://kg.virtualtreasury.ie/> .
2 @prefix crm: <http://erlangen-crm.org/current/> .
3 @prefix vrti: <https://www.w3id.org/virtual-treasury/ontology#> .
4 @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
5 @prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
6
7 ### Person: Oscar Wilde
8 <person/Wilde_Oscar-Fingal-OFlahertie-Wills_c19/v1z53fj>
9   a crm:E21_Person ;
10   crm:P1_is_identified_by
11     <normalized-appellation-forename/Oscar-Fingal-OFlahertie-Wills> ,
12     <normalized-appellation-surname/Wilde> ,
13     <normalized-appellation-surname-forename/Wilde_Oscar-Fingal-OFlahertie-Wills> ;
14   vrti:VRTI_REF_ID "v1z53fj" ;
15   vrti:DIB_ID "a9036" .
16
17 ### Appellations
18 <normalized-appellation-forename/Oscar-Fingal-OFlahertie-Wills>
19   a crm:E41_Appellation ;
20   rdfs:label "Oscar Fingal O'Flahertie Wills" .
21
22 <normalized-appellation-surname/Wilde>
```

¹<https://docs.gaois.ie/en/data/logainm/v1.0/data>

```

23   a crm:E41_Appellation ;
24   rdfs:label "Wilde" .
25
26   <normalized-appellation-surname-forename/Wilde_Oscar-Fingal-OFlahertie-Wills>
27   a crm:E41_Appellation ;
28   rdfs:label "Wilde, Oscar Fingal O'Flahertie Wills" .
29
30   ### Birth event
31   <birth/Wilde_Oscar-Fingal-OFlahertie-Wills_c19/v1z53fj>
32   a crm:E67_Birth ;
33   crm:P98_brought_into_life
34       <person/Wilde_Oscar-Fingal-OFlahertie-Wills_c19/v1z53fj> ;
35   crm:P4_has_time-span
36       <time-span/1854-10-16_1854-10-16> ;
37   crm:P7_took_place_at
38       <place/present-day/county/Dublin/v1s4kv2> .
39
40   ### Birth time-span
41   <time-span/1854-10-16_1854-10-16>
42   a crm:E52_Time-Span ;
43   crm:P82a_begin_of_the_begin "1854-10-16"^^xsd:date ;
44   crm:P82b_end_of_the_end "1854-10-16"^^xsd:date .
45
46   ### Birth place
47   <place/present-day/county/Dublin/v1s4kv2>
48   a crm:E53_Place ;
49   crm:P1_is_identified_by
50       <appellation/place/Dublin-en> ,
51       <appellation/place/Dublin-ga> .
52
53   <appellation/place/Dublin-en>
54   a crm:E41_Appellation ;
55   rdfs:label "Dublin"@en .
56
57   <appellation/place/Dublin-ga>
58   a crm:E41_Appellation ;
59   rdfs:label "Baile Átha Cliath"@ga .

```

The combination of event-mediated modelling, high abstraction, and multi-ontology integration results in a complex KG in which relationships between entities are distributed across multiple classes and properties. Understanding how the information is structured or how entities are connected presents a challenge for lay users, particularly lay users encountering the VRTI KG for the first time. Such characteristics directly contribute to the conditions that give rise to the Initial Exploration Problem (IEP) framing presented by this thesis.

2.3 Digital Humanities KG User Interfaces (UIs)

As outlined in the previous section, cultural heritage KGs such as the VRTI KG are typically modelled using CIDOC CRM and related ontologies. While use of such ontologies

enables rich semantic representation and interoperability, it also produces structurally complex KGs in which relationships are mediated through events. As a result, direct interaction with a KG that utilises such ontologies is infeasible for most lay users. Instead, interaction with these KGs is primarily carried out through user interfaces. These interfaces typically support searching, browsing (including faceted browsing) as discussed in the next section.

2.3.1 Existing KG-Based Historical UIs

A number of interface frameworks and project-specific implementations have emerged to support the exploration of humanities KGs. Among the most widely used is Sampo-UI [20], developed by the Finnish Semantic Computing Research Group (SeCo). Sampo-UI is a configurable web application framework designed specifically for cultural heritage and historical KGs. It has been deployed across multiple "Sampo" portals, including LetterSampo [38] (Figure 2.1), which provides access to early modern correspondence networks. Sampo-UI is structured around the concept of perspectives, allowing users to view the same underlying data through different coordinated visualisations such as tables, maps, timelines, and network graphs. These perspectives are powered by SPARQL query templates connected to an underlying endpoint. This approach enables rich multi-perspective exploration of historical data but assumes that users can already identify meaningful entry points into the dataset.

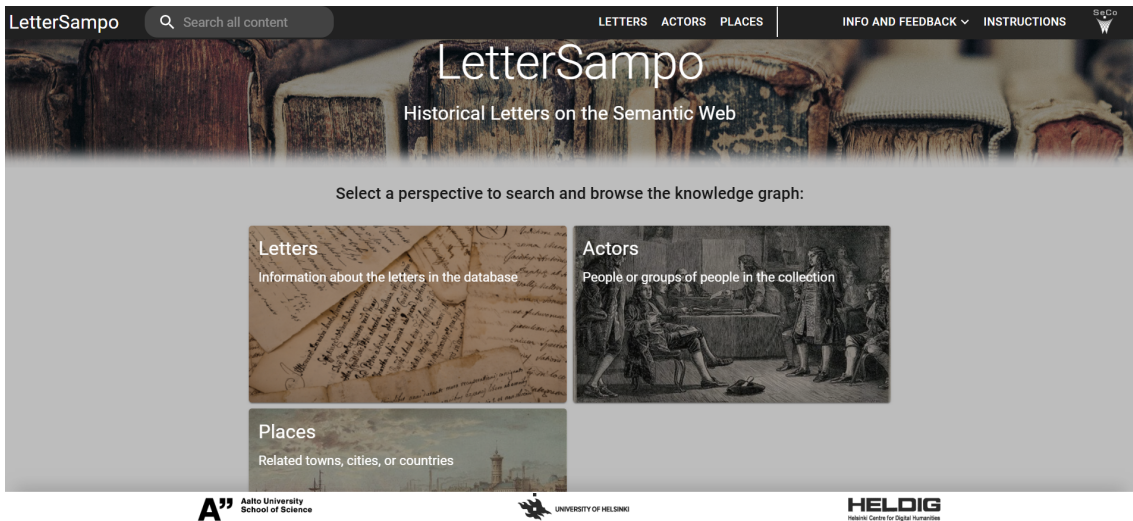


Figure 2.1: Homepage of the LetterSampo UI.

The Virtual Record Treasury of Ireland (VRTI) Explorer [37] provides another example of a digital humanities KG UI custom built for the VRTI KG. The VRTI Explorer is the UI used as part of the research undertaken during this thesis to investigate Curated Questions and Answers (CuQAs) as starting points to begin exploration. The VRTI Explorer enables users to search and browse the information contained within the VRTI KG and supports keyword search and perspective-based and entity-based exploration across persons, places, and records, presenting results through structured entity pages and linked navigation pathways. The UI is under active development and Figure 2.2 represents how the UI appears as of February 2026. As is explained in further detail in Chapter 5, the

interface underwent significant development during the course of the research with the addition of new features and changes in presentation. In addition to the VRTI Explorer, the underlying VRTI KG is accessible via a SPARQL endpoint, allowing direct querying through tools such as Virtuoso [58], HTML representation of the RDF via LodView [59], and visual representation of the ontology using Ontodia [60]. These UIs support both lay users and technically proficient ones; however, as with many humanities KG interfaces, the system primarily supports exploration once a user has identified a search term or entity of interest, providing limited guidance for users encountering the dataset for the first time.



Figure 2.2: Homepage of the VRTI Explorer.

Taken together, the modelling characteristics of CIDOC CRM-based KGs and the design of existing digital humanities exploration UIs reveal a structural gap between data representation and the onset of exploration by a lay user unfamiliar with the KG. These existing KG and linked data interfaces often prioritise search, querying, and visualisation functionality but provide limited support for helping unfamiliar lay users understand dataset scope or identify meaningful starting points for exploration, leaving many semantic datasets underused by their intended audiences [8, 61]. This gap between available functionality (interaction primitives) and initial user orientation motivated the framing of the Initial Exploration Problem (IEP) which is presented in depth in the next chapter.

Chapter 3

The Initial Exploration Problem (IEP) and Related Work

This chapter first presents a conceptual gap in Human-Computer Interaction (HCI), Semantic Web, and Knowledge Graph (KG) usability literature, where the challenges experienced at the onset of exploration of a large, complex KG have been addressed in a fragmented manner. It then introduces a conceptual framing for this "first-contact" problem, referred to in this thesis as the Initial Exploration Problem (IEP), and presented in support of this research's Minor Contribution 1 outlined in Chapter 1, Section 1.0.6. Finally, the chapter examines existing KG exploration interface design primitives through the lens of the IEP framing, highlighting the gap that is addressed through the introduction of Tús Maith and Curated Questions and Answers (CuQAs) in later chapters and the potential for hybrid interface solutions. In defining IEP and characterising it through a conceptual framing, this thesis argues that both clearer evaluation criteria and the development of targeted scaffolding techniques (e.g., curated questions and answers(CuQAs)) are enabled, not only in support of the research of this thesis but in support of researchers and practitioners in the community.

Throughout this thesis, the term *IEP* is used in two closely related but analytically distinct ways. Firstly, it names a recurring empirical problem observed when lay users encounter an unfamiliar, complex KG and struggle to identify how or where to begin exploration. Secondly, this chapter develops a conceptual framing of the IEP that explains the problem by characterising the conditions under which it arises. Where necessary, this thesis distinguishes between the IEP as a named problem and the conceptual framing of the IEP composed of specific, interdependent characteristics.

3.1 The Theoretical Underpinnings of the Initial Exploration Problem (IEP)

Whilst the barriers arising from the IEP are closely linked to the difficulties associated with lay users interacting with KGs, such difficulties can be better understood through looking at existing HCI and information behaviour concepts and theories. This section examines

these concepts and theories in an effort to show why no one existing term fully captures the IEP and sheds light on how the IEP has previously been explored in fragments across the literature.

3.1.1 Foundations in Information Behaviour: Uncertainty and ASK

At the onset of user interaction, exploration is fundamentally different from traditional information retrieval (IR) [62, 63]. Traditional IR typically assumes that the user has an information need that they are seeking to fill through asking a question over information [64]. Exploration in contrast does not assume such a need; the user instead is making their way through the data in what could be described, at least at first, as an open-ended fashion driven by curiosity as opposed to explicit questions. Through their exploration, questions may occur to the user at which point traditional IR comes into effect. More specifically, once on a path, the user can make decisions on where to go next, traversing the web of knowledge provided to them. The IEP specifically refers to the empirical difficulty observed at the onset of exploration, where curiosity is present but a goal is absent. The IEP framing characterises the difficulty of identifying a starting point for traversal, and the ways in which this difficulty is exacerbated by the format in which knowledge is represented (i.e., KGs).

The state of uncertainty of where to begin is analogous to Belkin’s Anomalous State of Knowledge (ASK) [65] where he identifies users looking at IR from a place of "perceived wrongness". Essentially, that the user’s current internal picture of their knowledge is flawed and they are looking to correct it, therefore mapping the problem to IR. The difference captured by the IEP framing is that the user’s sense of uncertainty does not come from a gap or flaw in their internal knowledge that they are looking to fix but instead, the uncertainty comes from a combination of challenges, namely the characteristics that make up the IEP: the user does not know the scope of the KG, they do not know how the data is structured, and they do not have the relevant expertise in using SPARQL in order to orient themselves. ASK presupposes a knowledge gap; the IEP precedes the ability to articulate such a gap. Where ASK assumes a misalignment between what a user knows and what they need to know, the IEP describes a prior state in which neither the nature of the gap nor the means of resolving it are yet apparent.

3.1.2 Frameworks for Process: Exploratory Search and Sensemaking

Exploratory search is an umbrella term under which KG exploration sits and is encountered here as exploration preceding information need. More specifically, KG exploration can be understood as a specific instantiation of exploratory search, characterised by explicit semantic structure and graph-based interaction. Marchionini [66] captures exploratory search as a combination of learning and investigating but does not isolate a user’s first contact with data; the assumption is that the user can begin "somewhere" and tools then should support their exploration journey. The problem with this is that when a lay user has no underlying mental model of the structure of the data or what they can ask, such a starting point is not obvious. A similar limitation arises with Russell et al.’s

sensemaking paradigm [29]; in this instance, there is the presence of a goal or "task", something which is at odds with exploration driven by curiosity. Similarly, while White et al. [67] acknowledge exploratory search can be driven by curiosity, in the absence of a goal, they do not investigate or propose a starting point from which curiosity-driven exploration can begin.

3.1.3 Cognitive Foundations: Load, Foraging, and Scaffolding

Information Foraging Theory [68] provides a complementary lens, modelling information seekers as predators following "information scent". The IEP represents a complete absence of scent; without prior knowledge of the KG's scope or structure, users cannot evaluate the utility of potential paths, leaving them "stranded" at the entry point with no trajectory to follow. This absence of information scent has direct implications for cognitive load. At first contact, users must simultaneously interpret unfamiliar terminology, infer structural conventions, and evaluate potential entry points, resulting in high intrinsic and extraneous cognitive load [69]. Unlike later stages of exploration, this load cannot be mitigated through strategy or experience.

3.1.4 Where do we Start? Orientation, Onboarding, and Serendipitous Discovery

When considering the issue of "where to begin" as it relates to human interaction with digitised data, the HCI concepts of orientation and onboarding are surfaced, referring to situations where users need initial guidance to build situational awareness in information systems. Onboarding in User Interface (UI) design emphasises progressive disclosure, greeting users with simple tours or anchors to reduce abandonment. In KG contexts, this could involve default starting entities or visual maps. However, onboarding is a method of solution to the IEP, it does not articulate the problem. In short, onboarding describes how systems introduce users to interfaces; the IEP describes the epistemic condition that makes such introduction necessary in the first place. Serendipitous discovery complements onboarding: while IEP blocks entry, tools fostering unexpected finds (e.g., following paths [70]) can spark exploration once initiated [71].

3.1.5 Synthesis

These underpinnings can be synthesised into a cohesive, clearly scoped account of the problem by isolating a distinct, first-contact condition in KG interaction: it precedes ASK by embodying pre-gap uncertainty, extends exploratory search by isolating the entry barrier, and leverages cognitive theories for solutions like scaffolded entry-points. Unlike sense-making's task-focus, the IEP framing highlights curiosity-driven starts in semantic spaces. Table 3.1 provides a comparison of where the IEP framing sits in relation to cognitive and usability literature. Although the concepts listed in the table originate from different theoretical traditions (ranging from cognitive psychology to interaction design), the purpose of listing the concepts is not to equate them analytically, but to position them temporally

within the user interaction lifecycle. Specifically, the table examines whether each concept presupposes that a user has already succeeded in initiating meaningful interaction with a KG.

Building on these theoretical fragments, the next section demonstrates how the Initial Exploration Problem manifests as a unified, observable phenomenon in KG contexts.

3.2 The Manifestation of the Initial Exploration Problem (IEP) in KGs

This section demonstrates that the problem referred to as the IEP is a real, observable phenomenon in the KG exploration domain by analysing the causes identified through the IEP framing. In the first three subsections, it illustrates how the IEP manifests in practice through recurring epistemic and interactional barriers observed in KG exploration research. Informed by these barriers, the next two subsections proceed to present this thesis author’s proposed framing of Initial Exploration Problem (IEP) in terms of three inter-dependent barriers with respect to interaction of users with KGs.

3.2.1 Complex Ontologies and Technicality

At first contact, complex ontologies do not merely impede query formulation; they prevent lay users from forming even a coarse understanding of what the KG contains, and therefore from identifying a plausible point of entry. Prior work such as [73] recognises the complications incurred in beginning to explore an unfamiliar KG by a lay person: the user’s mental model of the underlying structure of the knowledge does not match the true complex structure of the ontology. This leads to overwhelm and an inability to query effectively: "users do not know what they do not know" [65]. Moreover, there is an assumption that users understand sufficiently the domain in which they are beginning to explore to, if provided with facets or entities, be able to choose ones that will provide them with meaningful exploration starting points [74]. An assumption that does not hold in practice and constitutes as a direct manifestation of **scope uncertainty**. Furthermore, without an understanding of the ontology or conceptual schema underlying the KG (**ontology opacity**), lay users are unable to anticipate how information is organised or what forms of questions are meaningful. This disconnect between the user’s mental model and the system’s semantic structure has been repeatedly identified as a barrier to effective KG interaction [18] and as such, **ontology opacity** manifests not only as a difficulty executing actions, but as a difficulty anticipating what kinds of questions and relationships are meaningful within the KG.

It has long been established within KG usability literature that there is a significant barrier between lay users and the technical query language SPARQL [19]. This has led to numerous solutions aiming to hide the complexities of SPARQL behind more intuitive interfaces [75, 30]. However, such solutions do not look at the moment of first contact; they become effective only after the user knows in some capacity what it is that they are looking for. At first contact, **scope uncertainty** precedes **query incapacity**: users struggle not only

Table 3.1: A comparison of the Initial Exploration Problem (IEP) framing with related concepts in information seeking and interaction.

Concept	Primary focus	Assumes a starting point?	Assumes a goal or question?	Applies at first contact?
Exploratory search	Learning, investigation, discovery	Yes	Not necessarily	No
Sensemaking	Constructing understanding from information	Yes	Yes (explicit or implicit)	No
Information foraging	Navigational and information-seeking strategies	Yes	Implicitly	No
Berrypicking [72]	Iterative query refinement	Yes	Yes (evolving)	No
ASK (Anomalous State of Knowledge)	Knowledge gap triggering search	Yes	Yes	No
KG usability issues	Interaction and system-level barriers	Yes	Often	Partially
Onboarding	Design strategies for initial system use	Yes	No	Sometimes
Serendipitous Discovery	Unexpected and valuable discoveries during exploration	Yes	No	Sometimes
Information overload	Cognitive effect of excessive information	Yes	Not required	Sometimes
Initial Exploration Problem (IEP)	Entry into exploration of an unfamiliar KG	No	No	Yes

to express queries, but to determine what kinds of questions are possible to be asked in the first place.

3.2.2 Scale and Interconnectedness

Scale and interconnectedness do not introduce new barriers, but intensify the epistemic opacity faced at first contact, making it harder for users to infer what constitutes a promising direction of exploration. For instance, large KGs like DBpedia [6], with billions of triples, exemplify how the problem is exacerbated by presenting a hyper-connected space where every entity links to dozens or hundreds of others, leading to confusion and high cognitive load [8]. This is further discussed by Al-Tawil et al. [73] with the authors noting that "users can face an overwhelming amount of exploration options and may not be able to identify which exploration paths are most useful" which "can lead to confusion, cognitive load, frustration, and feeling of being lost". This results in a profound difficulty of understanding the scope (**scope uncertainty**) of the KG by lay users.

Furthermore, while lay users may enter with vague preconceptions or hypotheses about potential content (e.g., assuming a historical KG contains information on well-known historical events), these assumptions can mislead: what they believe might exist may not (leading to fruitless searches), or if it does, it may represent a peripheral or isolated subset of the KG, resulting in exploratory "dead ends" obscuring the true breadth and interconnectedness of the KG. While the IEP is most visible in curiosity-driven exploration, it also applies when lay users possess only vague or underspecified goals, insofar as they lack the knowledge required to identify a meaningful entry point into the KG.

3.2.3 The IEP in the Digital Humanities (DH) Domain

The IEP presents as a manifestation of a combination of domain-agnostic problems compounded by the added difficulties experienced by lay people in the Digital Humanities (DH) domain. This is due to the often complex ontologies used (e.g. CIDOC CRM [12]), the scale of the KGs in question (e.g. the VRTI KG at 2.9 million triples [10]), and the persistent difficulty of the query language SPARQL [16] for lay users. Studying the IEP in the DH is therefore an ideal use case: the KGs themselves are rich and expressive, but the IEP may hinder their exploration by lay users. For lay users, these characteristics compound **scope uncertainty** and **ontology opacity**: not only is the domain unfamiliar, but the intellectual organisation of the knowledge itself is opaque, making it difficult to infer what kinds of historical questions the KG can meaningfully support.

3.2.4 The Characteristics of the IEP

Within the proposed framing, the IEP is characterised as the co-occurrence of three interdependent barriers that arise simultaneously at first contact and can be distilled as:

Scope Uncertainty: Scope uncertainty represents the core problem of "where to begin". Where do you start if you have no knowledge of what is contained within the KG? This is exacerbated in large-scale KGs where users cannot gauge the breadth of topics, leading

to cognitive overload [8] and is therefore epistemic rather than technical: it concerns the user’s inability to reason about the knowledge space itself.

Ontology Opacity: The structured nature of a KG promotes interoperability which is one of the reasons as to why they are so valuable for knowledge representation. However, this structure can often be difficult for lay users to interpret and map mentally [76], particularly in the absence of exemplar queries or conceptual anchors.

Query Incapacity: SPARQL as a query language has long been acknowledged as a significant barrier to lay people utilising KGs [77, 78, 76, 19].

3.2.5 The IEP Framing as a Unified Concept

Across these manifestations, scope uncertainty emerges as the dominant first-contact breakdown: ontology opacity and query incapacity become problematic precisely because users lack an initial understanding of what the KG contains and how exploration might productively begin. Even so, the characteristics within the IEP framing do not exist in isolation, all are present at the onset of exploration of an unfamiliar, large, complex KG by a lay user. Addressing one does not alleviate the other two; for example, providing users with a means of asking NLQs over a KG does not address the problem of what to ask (**Scope Uncertainty**) or, how to structure the question (**Ontology Opacity**). Similarly, in knowing what one wants to ask, one does not necessarily have the ability to structure the question in a manner that aligns with the underlying structure of the KG (**Ontology Opacity**) or have the necessary technical skill to create SPARQL queries (**Query Incapacity**). It is therefore important to consider the problem holistically when designing solutions that aim to address it.

3.3 The IEP as a Constraint on KG Interface Design

The IEP is temporally-bounded: it presents itself at the very start of KG exploration. It is because the problem is temporally-bounded that addressing it requires targeted solutions distinct from those addressing ongoing exploration. This thesis argues that addressing the first contact challenge in KG exploration in isolation is necessary as it is one that, unless considered as a problem in its own right, may lead to solutions developed that do not adequately address the barriers encountered. Crucially, this difficulty is not limited to identifying an interface action, but reflects an inability to determine what constitutes a meaningful or valid starting point within the knowledge space itself. In particular, it suggests that many existing interfaces fail to address the IEP not because of shortcomings in implementation, but because of the fundamental design assumptions embedded in their interaction models.

Rather than examining individual systems as isolated artefacts, this section analyses KG exploration interfaces at the level of interaction primitives and epistemic dependencies. That is, it asks what kinds of user knowledge an interface assumes must already be in place in order for interaction to begin. This shift in unit of analysis allows the IEP framing to function as a design constraint on the space of possible interface solutions, rather than

as an evaluative checklist applied post hoc.

3.3.1 Solution Development Pre-Problem Formulation

In the broader traditions of Human-Computer Interaction (HCI) and Semantic Web research, it is not uncommon for practical systems to be developed in response to recurring user difficulties before those difficulties are formally conceptualised as distinct research problems. This phenomenon, where design precedes theoretical articulation, has been noted in several areas of computing and information science. For instance, the emergence of recommender systems in the late 1990s and early 2000s [26] responded to the growing practical problem of information overload on the web. Only in subsequent years was the challenge rigorously theorised within information behaviour research as a cognitive and affective phenomenon [79]. Similarly, in NLQ interfaces to databases and KGs, early tools such as START [80] and more recent QA systems like PowerAqua [81], as well as later QA evaluation campaigns such as QALD [82], addressed usability barriers long before formal taxonomies of query intent [83], exploration vs. lookup [66, 67, 64], or onboarding challenges [8] in KG interaction were widely discussed in academic literature.

Naming and characterising the IEP is further motivated by the development of systems such as OnSET [28] and Knowledge Anchors [73]. Such systems can be understood as independently developed responses to a shared but unnamed challenge: enabling users to initiate meaningful exploration of complex KGs. While these approaches differ in modality (visual, conceptual, and cognitive), they each introduce scaffolding intended to overcome the same initial barrier faced by lay users. Upon closer inspection, however, a gap in conceptualisation becomes evident. Al-Tawil et al.’s [73] approach of Knowledge Anchors frames KG interaction as a learning process leaning on subsumption theory [84], the idea that people learn better by associating new knowledge with the knowledge they already possess. This theory however assumes that a user can select a meaningful starting point with which to build from. The IEP framing makes this assumption explicit and problematises it, thereby providing the missing theoretical foundation for such practical interventions.

3.3.2 Interaction Primitives and Epistemic Preconditions

It is possible, drawing on the work of [85], [86], and [87], to decompose any KG exploration interface into a small number of interaction primitives such as searching, filtering, selecting, browsing, or querying. While these primitives differ in modality (textual, visual, graphical), they share a common property: each presupposes certain epistemic conditions on the part of the user. For example, issuing a search query presupposes that the user has and can articulate a meaningful information need, while selecting a class or facet presupposes some awareness of the underlying schema or domain structure and what those classes or facets represent.

These assumptions extend beyond interaction mechanics to the epistemic role of exploration interfaces themselves. In KG environments, interface representations do not simply expose data but shape how users form an understanding of the knowledge space. Visualisations, schema views, and graph navigation structures implicitly communicate what

entities, relationships, and questions are salient within the KG. Recent work in knowledge organisation highlights that KG visualisation systems often prioritise structural or algorithmic representation while paying less attention to how users cognitively interpret those representations [88]. Consequently, interface design influences not only how users interact with a KG but also how they construct a mental model of what the KG contains and what kinds of questions it can meaningfully support.

Within the IEP framing, these presuppositions become problematic. At first contact, a defining characteristic of the IEP is precisely the absence of such epistemic conditions: the user does not yet know what the KG contains (scope uncertainty), how it is structured (ontology opacity), or how to express questions in a formal language (query incapacity). Interaction primitives that rely on any of these forms of prior knowledge are therefore structurally misaligned with the moment at which the IEP arises.

Table 3.2 summarises common interaction primitives used in KG exploration interfaces and the epistemic assumptions they impose.

Table 3.2: Interaction primitives in KG exploration interfaces and their epistemic preconditions.

Interaction primitive	Primary function	Epistemic precondition
Keyword or entity search	Locate relevant entities or facts	User has and can articulate a meaningful query or term
Facet selection	Narrow result space	User understands facet semantics and relevance
Schema or class browsing	Reveal ontology structure	User can interpret abstract ontological concepts
Graph navigation	Traverse linked entities	User has an initial conceptual anchor
Visual query construction	Build structured queries	User understands relationships and constraints
NLQA	Retrieve answers via natural language	User can articulate an information need in natural language

Viewed through this lens, it becomes apparent that most existing KG exploration interfaces are designed around primitives that presuppose the successful resolution of at least one of the IEP characteristics. Consequently, they are effective only after the IEP has been overcome, either through prior knowledge or external guidance.

3.3.3 Paradigms of KG Exploration as Compositions of Primitives

Existing KG exploration systems can be understood as compositions of the interaction primitives outlined above. Faceted browsers (e.g., KG Explorer [89]) combine search and filtering; visual exploration tools (e.g., Ontodia [60], KGViz [24], and PG-Explorer [90]) combine schema exposure and graph navigation; natural language interfaces (e.g., SPARK-LIS [30]) combine question formulation with query translation. While these paradigms differ in surface interaction style, they are unified by a shared assumption: that the user

is already capable of selecting a meaningful starting point.

From the perspective of the IEP framing, this assumption is precisely what fails at first contact. Interfaces that prioritise navigation, refinement, or query formulation implicitly position the act of starting as trivial or self-evident. The IEP as a conceptual framing identifies this moment as the core challenge. As a result, these paradigms can be said to address usability within exploration, while remaining structurally incapable of initiating exploration under the epistemic conditions of the barriers of the IEP.

Importantly, this is not a claim that such systems are ineffective or poorly designed. Rather, it is a claim about temporal alignment: the problems they are designed to solve occur later in the interaction lifecycle than the IEP itself. This explains why multiple, independently developed systems across the Semantic Web and Digital Humanities communities converge on similar limitations when deployed for lay users encountering a KG for the first time (e.g., SPARKLIS, PG-Explorer, KGViz, Sampo UI [20]).

3.3.4 The Missing Interaction Primitive: Scope Revelation

The analysis above reveals a structural gap in the design space of KG exploration interfaces. While existing paradigms provide strong support for query formulation, schema inspection, and exploratory navigation, they lack an interaction primitive whose primary epistemic function is scope revelation without user specification. That is, there is no dominant design pattern whose purpose is to communicate what the KG contains and what kinds of questions it can answer before the user is required to act.

Within the IEP framing, such a primitive must satisfy three conditions. First, it must not require the user to articulate a goal or query. Second, it must not rely on the user interpreting ontological structures or terminology. Third, it must function as a meaningful conceptual foothold from which further exploration can proceed. Existing primitives fail to meet these conditions simultaneously, explaining why scope uncertainty persists even in interfaces that successfully mitigate ontology opacity or query incapacity.

This observation reframes the problem of KG exploration at first contact as one of entry-point design. The absence of an appropriate entry-point primitive is not accidental but reflects a broader tendency in interface design to privilege interaction over orientation as surfaced in Section 3.1. Addressing the IEP therefore requires not incremental improvements to existing paradigms, but the introduction of interaction constructs explicitly designed for epistemic onboarding.

3.3.5 Implications for IEP-Oriented Interface Design

By treating the IEP and its framing as a design constraint rather than a usability flaw, it becomes possible to articulate clearer criteria for what constitutes an IEP-oriented interface. Such interfaces must incorporate interaction primitives whose function is not to respond to user intent, but to help form that intent by revealing the scope and structure of the KG in cognitively accessible terms, in other words, orienting the user.

This perspective also clarifies the role of hybrid interface solutions. Interfaces that combine traditional exploration mechanisms with explicit entry-point scaffolding can support both

initiation and sustained exploration, provided that these components are temporally and conceptually separated. The following chapters build on this insight by introducing the Curated Question and Answer (CuQA) as a concrete instantiation of a scope-revealing interaction primitive that also addresses ontology opacity and query incapacity, and by presenting the Tús Maith framework as a means of supporting their creation and curation at scale.

3.4 Summary and Implications

This chapter has introduced and defined the Initial Exploration Problem (IEP), and developed an IEP framing that theorises it as a distinct, critical barrier to lay user orientation at the precise moment of first contact with an unfamiliar, complex KG, a problem previously only addressed fragmentarily in both literature and in system design. In naming and characterising it through the IEP framing, this chapter provides the theoretical basis for the design and evaluation of scaffolding mechanisms introduced in subsequent chapters. Through the IEP framing, the problem has been distilled into three interdependent characteristics: *Scope Uncertainty*, *Ontology Opacity*, and *Query Incapacity*. By examining the theoretical basis from which the problem arises, the IEP framing positions it within, yet distinct from, established concepts like ASK, exploratory search, and cognitive load theory, as summarised in Table 3.1. Its uniqueness lies in its temporality (applying strictly at first contact) and its pre-conditional nature (assuming no starting point or explicit goal). The contribution of the IEP framing is therefore not the identification of new individual barriers, but the isolation and characterisation of their convergence at a specific temporal point in interaction: the moment of first contact. While the IEP is most visible in curiosity-driven exploration, it also applies when lay users possess only vague or underspecified goals, insofar as they lack the knowledge required to identify a meaningful entry point into the KG.

Defining the IEP and developing its conceptual framing has two primary implications corresponding to the dual aims of Minor Contribution 1. Firstly, it establishes a clear conceptual lens and vocabulary through which to evaluate how systems support (or fail to support) the initial exploration phase. It is now possible to assess existing solutions and interface primitives based on how effectively they mitigate Scope Uncertainty, clarify Ontology Opacity, and bypass Query Incapacity. In doing so, it follows a long-standing trajectory in HCI: that of moving from ad hoc design responses to theory-informed interface paradigms, ultimately improving both usability and explanatory power. Secondly, it provides the foundation for the design of targeted scaffolding techniques. The subsequent chapters will introduce and evaluate one such technique: Curated Questions and Answers (CuQAs), which is the target artefact produced by the Tús Maith framework and its instantiation TMv1. By first isolating and naming the problem, this thesis moves from ad hoc interface design to theory-informed intervention, aiming to transform the initial moment of confusion into a structured and supported entry into exploration.

Chapter 4

The Design, Instantiation, and Implementation of Tús Maith

4.1 Chapter Objectives

This chapter provides the design basis for the Tús Maith framework grounded in the Curated Question and Answer (CuQA) lifecycle, describes its instantiation as the TMv1 system, and details the architectural and implementation choices that realise this design in practice. It directly addresses RQ2 by instantiating the Tús Maith framework as TMv1, supporting RO2 through detailed architectural and implementation choices. The chapter establishes the conceptual foundations of the framework, specifies its interaction model and epistemic lifecycle, and provides sufficient technical detail to support reproducibility and to contextualise the evaluations presented in the following chapter. The Tús Maith framework was developed from the need to be able to create Curated Questions and Answers (CuQAs) by non-technical domain experts, where CuQAs are NLQAs curated by domain experts to act as exploratory entry points into a KG thereby aiming to mitigate the Initial Exploration Problem (IEP). As discussed previously, the IEP is a temporally bounded one (a problem that occurs at the onset of exploration) and as such, the design of Tús Maith, the instantiation and implementation of TMv1, and the subsequent evaluations presented in Chapter 5 are focused solely on initial contact with an unfamiliar KG. While CuQAs may support continued exploration, the framework, instantiation, and evaluations presented here and the next chapter are explicitly concerned with the moment of first contact, as characterised by the IEP framing.

This chapter adopts a design-justified approach to literature, in which prior work is introduced at the point where it motivates or constrains a specific architectural, epistemic, or interactional design decision. Rather than presenting a standalone review, relevant literature is embedded within the rationale for the framework’s lifecycle structure, its human-in-the-loop governance model, and its entity-centric anchoring architecture.

4.2 The Tús Maith Framework

This section introduces the conceptual foundations of Tús Maith and motivates its decomposition into three distinct stages. The framework is presented independently of any particular system instantiation, with the intent of demonstrating its generality and applicability across different technical realisations.

Prior interfaces have aimed to make KGs more accessible to non-experts. Systems like OnSET [28] and SPARKLIS [30] provide visual query-building and schema exploration tools. Usability studies consistently report that lay users struggle to discover KG structure and formulate queries. Recent works (e.g. [91, 44, 19]) document this initial exploration challenge and propose guided query suggestions. This body of work motivates the use of pre-defined templates in Tús Maith as natural-language entry points, positioned specifically to support first contact rather than general-purpose query formulation.

4.2.1 Design Motivation: The Curated Question and Answer (CuQA) Lifecycle

The Tús Maith framework is derived from the articulation of the lifecycle of curated natural language questions and answers (CuQAs). CuQAs are intended to function as NLQA-based KG exploration entry points that reveal the scope of a KG, ground user interaction in its underlying structure, and remove the need for direct engagement with SPARQL. Such CuQAs can be considered to fit within and extend the "Visual Interfaces Built on Top of KGs" category proposed by Li et al.[15] as part of their KG interfaces that support organic discovery. While they distinguish between open-ended and goal-oriented exploration, they do not identify the onset of exploration as a distinct challenge and their proposed solutions (e.g., query-building interfaces) within that category require interaction by the user: something which requires a starting point (which is misaligned with the constraints of the IEP framing).

To support this role, the lifecycle is decomposed into three stages. First, abstract competency questions (CQs), traditionally used for ontology development, are repurposed as reusable, schema-level templates that define the semantic scope of potential interactions without reference to specific entities (Stage 1). Second, these templates are instantiated by binding them to concrete KG entities, grounding the resulting NLQA pairs in the contents of the graph (Stage 2). Finally, the instantiated artefacts are subject to domain expert review and curation, through which selected pairs are promoted to CuQAs suitable for deployment as exploration entry points (Stage 3).

4.2.1.1 CQ Templates as Exploration Starting Points

CQs were originally introduced as requirements-driving queries in ontology engineering [34]. The research described in this thesis repurposes them from that original purpose (to provide a means of validating and assessing ontologies), to instead, as templates, form a set of natural language questions (NLQs) that expose the scope of a KG in an ontologically aligned manner. This repurposing preserves the core property of CQs, namely, their role

as schema-aligned articulations of what a KG can express, while shifting their function from formal validation to communicative scaffolding. In this way, the latent nature of CQs (revealing the scope of a KG in natural language) is being leveraged to support the onset of exploration by lay users.

Well designed, reusable templates have been shown to support structured question formulation and reuse in KG question answering settings [92, 93]. In such systems, templates enable users to generate multiple concrete questions by binding entity placeholders within shared natural-language patterns. More broadly, abstraction mechanisms such as merging and clustering related information while suppressing less salient attributes have been shown to support interpretability and manage cognitive load in complex information environments [8]. While visualisation provides a rich design space for this form of abstraction, prior work also proposes that text-based representations can benefit from similar structuring principles [8].

4.2.1.2 Utilising CQ Templates vs. Direct NLQA Derivation

Directly deriving NLQAs from individual entities would limit their capacity to convey the broader semantic scope of the KG and would preclude their reuse as generalised entry points. In contrast, deriving CQs at the schema level decouples the formulation of exploratory questions from any single entity and avoids a one-to-one correspondence between NLQs and SPARQL queries. As discussed in Section 4.2.4, this decoupling enables questions to function as mediating artefacts over entity-level context rather than as retrieval instructions.

While the lifecycle is presented as a linear progression from schema-level abstraction to entity-level grounding and expert curation, it also admits the interpretation of templates and instantiations as structured traversals over the semantic topology of the KG, enabling both exploratory breadth and controlled, multi-step semantic depth (discussed further in Section 4.2.4 below).

4.2.2 Framework Definition

As presented in Figure 4.1, Tús Maith is defined as a three-stage framework for the curation of NLQA entry points: (1) the derivation of schema-driven CQ templates, (2) the instantiation of these templates with KG entities, and (3) the expert review and curation of resulting NLQAs to produce curated NLQAs (CuQAs). The framework specifies the stages, responsibilities, and control points for human decision-making involved in this lifecycle, while remaining agnostic to the particular techniques used to realise each stage. Furthermore, the framework also remains agnostic to the degree of semantic depth or traversal complexity employed during CQ template instantiation, allowing both single-hop and multi-hop exploration strategies to be accommodated within the same lifecycle.

While the final curation stage constitutes a mandatory checkpoint requiring expert judgement, earlier stages may be instantiated using human, automated, or hybrid approaches. This separation of lifecycle structure from technical realisation means that Tús Maith has the potential to be applied across a range of domains and system architectures without

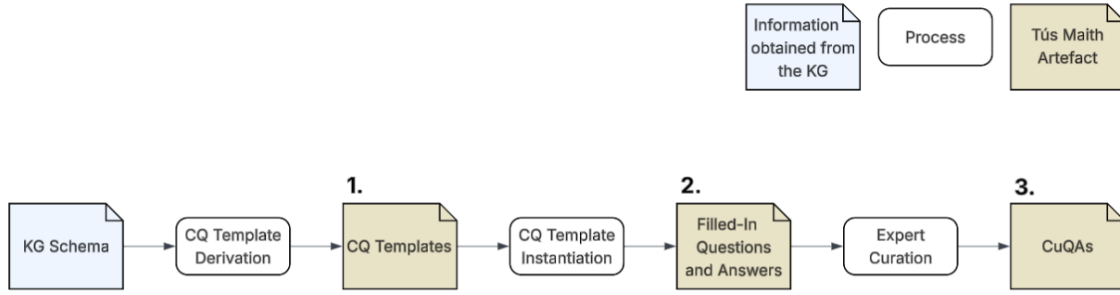


Figure 4.1: The Tús Maith framework: A framework to create CuQAs.

constraining the choice of underlying methods.

4.2.3 Human Judgement

Human judgement is an integral and non-optional component of the Tús Maith framework. The task of curating NLQA entry points is inherently epistemic and interpretive, requiring decisions about relevance, appropriateness, and communicative clarity that cannot be fully specified through formal rules or automated metrics. Because these entry points function as communicative artefacts intended to guide human exploration of a humanities KG, their quality must be assessed in relation to shared human understanding rather than computational correctness alone. While automated techniques may assist in the generation or filtering of candidate questions, the act of curation entails responsibility for the knowledge presented and therefore necessitates expert human judgement. As such, Tús Maith is human-in-the-loop by necessity, rather than by design preference.

This positioning aligns with work on "generous interfaces" in cultural heritage systems, which emphasises showing, contextualising, and orienting users rather than requiring them to articulate formal queries [94, 70]. In this way, the framing is consistent with interpretive and sensemaking perspectives in digital humanities, which treat system outputs as situated representations rather than objective answers. Furthermore, cognitive research on exploratory search systems in the digital humanities highlights that when systems mediate interpretations rather than strict retrieval, responsibility for validity and meaning must remain with human decision-makers [95]. Goodale et al. [95] apply this concept via Personalised Access To cultural Historical Spaces (PATHS) which align with CuQAs as structured domain expert curated navigation entry points.

More concretely, different instantiations of the framework may employ human, rule-based, LLM-supported, or hybrid techniques to execute generative functions. However, domain experts must retain decision authority throughout the lifecycle, including the selection of CQ templates, the choice of entities used for instantiation, and the criteria by which filled-in questions and answers are promoted to CuQAs.

4.2.4 Entity-Centric Query Anchoring Model

Tús Maith adopts an entity-centric query anchoring model that distinguishes exploratory interaction from question-to-query translation paradigms commonly employed in information retrieval and KG question answering (KGQA) systems. Rather than translating NLQs into query structures, CuQAs in Tús Maith are bound to specific KG entities, and user interaction triggers a fixed, schema-driven query that materialises the semantic neighbourhood of the anchored entity (or entities).

Within this model, CQ templates are derived solely from the schema-level classes and properties of the KG, without reference to SPARQL or to particular entities. To ground a template, a schema-driven entity materialisation query is executed for a selected entity or entities, retrieving their local graph context and serialising the result into a structured representation (e.g., JSON). This representation provides the contextual substrate from which the template is instantiated, producing a filled-in NLQ and a corresponding answer bound to the same entity or entities.

For example, a CuQA `Where was Oscar Wilde born?` `Dublin, Ireland` is tied to the URI for `Oscar Wilde`. When a user follows the question to the answer, a general query is executed over the KG using the URI to return any information in the immediate semantic neighbourhood about `Oscar Wilde` e.g., his forename, surname, date of birth, place of birth, etc. This ensures that when the user is presented with `Dublin, Ireland` as the curated answer it is not just a "dead end", there is further contextual information about `Oscar Wilde` present that they can investigate. Thus, the CuQA is an exploration starting point as opposed to a strict information retrieval mechanism.

When promoted from a filled-in question and answer to a CuQA through expert curation, a CuQA is stored as a tuple comprising of the CuQA and an entity URI. Upon user interaction, the same entity materialisation query is re-executed, presenting the broader graph context of the entity while foregrounding the curated answer as a particular semantic perspective over that context. In this way, the query functions as a mechanism for context materialisation rather than for question answering, positioning CuQAs as exploration primitives anchored in entity-level views of the KG.

In this entity-centric anchoring model, retrieval is performed as a bounded, schema-driven materialisation of an entity's local subgraph, without the specification of admissible semantic or topological paths. Semantic structure is introduced only at the stage of template binding and expert curation, where selected relations are projected as meaningful relational perspectives over an otherwise exhaustive entity context. This positions CuQAs as curated semantic frames rather than as the result of explicit graph traversal or path-based reasoning.

This prioritisation of entity-level context aligns with prior work showing that structurally dense, highly connected entities serve as effective entry points for sensemaking in large KGs, particularly when drawn from upper levels of the class hierarchy [74]. In Tús Maith, this principle informs the entity ranking strategy employed during CQ template instantiation (Appendix A.6).

4.3 TMv1: Instantiating the Tús Maith Framework

TMv1 is a system that instantiates the Tús Maith framework, employing large language models (LLMs) to support the first two stages of the lifecycle: schema-driven CQ template derivation and entity-based CQ template instantiation. Owing to the stochastic nature of LLM-supported generation, TMv1 adopts a conservative control allocation strategy that introduces additional human oversight during template selection and entity instantiation. TMv1 with human oversight is presented in Figure 4.2.

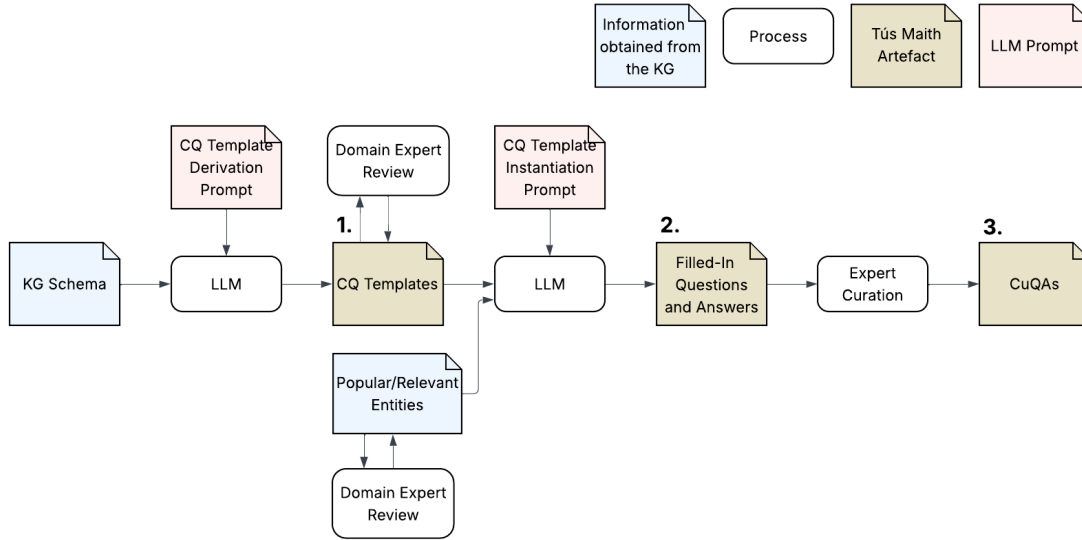


Figure 4.2: TMv1: An LLM-supported instantiation of the Tús Maith framework.

This oversight reflects properties of the chosen system realisation rather than requirements imposed by the framework itself. In principle, the Tús Maith framework admits alternative instantiations that employ rule-based, fully human, or other automated techniques, provided that the epistemic commitments and lifecycle structure specified in Section 4.2 are preserved.

4.3.1 Rationale for LLM-Supported Stages

The rise of LLMs has provided an alternative to rule-based or human-based knowledge acquisition, derivation, and instantiation methods. LLMs offer scalability but with trade-offs in cost/latency, as examined in Chapter 5. They are readily available and flexible, which makes them ideal candidates (with appropriate caveats, for instance, accounting for their propensity to hallucinate) to perform in Stage 1 and Stage 2. An expanded version of the table seen in Chapter 1 (Table 1.1) is presented below in Table 4.1 that describes the trade-offs between the various methods of instantiation.

Table 4.1: Comparison of Methods of Instantiation of Tús Maith.

Dimension	Human	Programmatic	Large Language Model (LLM)
Cognitive & Labour Cost	Cognitively demanding and time-intensive for curators	Low cognitive cost once implemented	Low human labour once prompt structure stabilised
Scalability	Not scalable beyond small expert teams	Limited scalability and new templates must be hand-coded	Highly scalable and can generate large, diverse question sets rapidly
Flexibility / Coverage	Experts can reason creatively but are bounded by familiarity and time	Brittle, coverage is restricted to predefined templates or rules	High coverage; can generalise to unseen schema patterns and phrasings
Naturalness of Language	Domain-appropriate	Often rigid or unnatural phrasing	High fluency
Transparency / Explainability	Fully transparent; reasoning traceable	Transparent logic but limited semantic nuance	Opaque generative process; requires provenance logging and human review
Trustworthiness / Accuracy	High, subject-matter expertise	High when rule set validated	Variable, may hallucinate or over-generalise; requires expert verification
Maintenance / Adaptability	Easily adaptable but costly in time	Hard-coded; requires developer effort to modify	Adaptable through prompt tuning or fine-tuning
Reproducibility / Provenance	Easy to track authorship	Fully reproducible if deterministic	Must record model, version, prompt, temperature to ensure reproducibility
Alignment with Tús Maith	High quality but slow	Systematic but inflexible	Balanced trade-off: scalable, expressive, domain-tunable, but requires oversight

For the instantiation of Tús Maith in TMv1, LLMs represented the method most aligned

with the purpose of TMv1: namely, a reduction in the need for technical knowledge when compared to a rule-based approach, being scalable, reducing the cognitive demand required by domain experts, and providing natural-language output. While their stochastic nature affords them a level of creativity and the ability to act on unseen data (unlike rule-based approaches), this nature introduces a higher requirement for human oversight throughout the stages of Tús Maith.

4.3.2 Oversight and Control Strategy

In utilising LLMs as the method of instantiation for Stage 1 and Stage 2 of TMv1, for the reasons outlined in the previous section, a greater need for human oversight and control was introduced. This need was realised through a human-in-the-loop LLM-supported system that enables domain expert decisions to dictate the curation of the CuQAs lifecycle end-to-end. In short, the LLMs are there only to act as cognitive aids to the domain experts, with the experts retaining all oversight.

More specifically, once CQ templates are derived from the classes and the properties of the KG, the domain expert has full control in deciding which of the templates will be instantiated. In Stage 2, while the domain expert is provided with popular search terms and the entities associated with those search terms, they have the ultimate control in deciding which entities specifically can be used to instantiate the CQ templates and, if they desire, to input entities of their own. Finally, once the CQ templates have been instantiated, producing filled-in questions and answers, the domain expert is the one who decides which of the filled-in questions and answers meets the criteria to become a CuQA and therefore can be presented to the end user.

While designed to be domain agnostic, TMv1 has been deployed using a digital humanities KG (the VRTI KG). The epistemic grounding of human oversight in its curation process aligns with broader calls in contemporary AI scholarship for systems that support collaborative human-AI decision-making. For example, a recent white paper report by the Alan Turing Institute, "Doing AI Differently - Rethinking the foundations of AI via the humanities", [96] argues that future AI systems should be designed to work with human interpretive capacities and cultural context, emphasising human-AI ensembles in which human judgement and expertise remain central to meaning-making and the responsible use of AI.

4.3.3 Stage 1: Existing CQ Template Derivation Systems and Approaches that Utilise LLMs

To the best of this thesis author's knowledge, no existing approach to date has derived CQs using LLMs for the purpose of providing exploratory entry points. Furthermore, no one existing method or experiment was immediately fit for purpose as part of TMv1 as presented below necessitating LLM Experiment 1 (Chapter 5, Section 5.7), which examined the extent to which LLMs can generate CQ templates to act as exploratory entry points. Recent work [48, 97, 98, 99] has utilised LLMs to automatically generate CQs alongside their associated SPARQL queries [100] from KGs.

More specifically, RevOnt [48] explores the reverse engineering of CQs from KGs such as Wikidata, drawing on a set of human-authored CQs from existing corpora as ground truth (GT) for evaluating output quality. Their definition of a CQ, "a typical query that an expert might want to submit to a knowledge base of its target domain, for a certain task", reflects a focus on ontology modelling and evaluation. In contrast, Tús Maith defines a CQ as a NLQ that helps convey the scope and potential of a KG to users unfamiliar with its structure, with the aim of supporting exploration rather than formal assessment. Moreover, RevOnt assumes the availability of triple verbalisations, a condition that does not hold for the VRTI KG, which contains over 2.9 million non-verbalised triples and therefore makes the RevOnt pipeline non-transferable to this context.

Other work [98] focuses on generating CQs from KGs constructed from PDF documents. After clustering similar entities and generating summaries, CQs are produced from these summaries and evaluated using zero- and few-shot prompting with cosine similarity scores. However, this introduces a degree of circularity by using the summary to both generate and evaluate the questions. Their work also took an LLM-as-a-judge evaluation approach without the involvement of domain experts, while in contrast, this thesis emphasises human expert evaluation to assess the quality, relevance, and answerability of generated CQ templates.

A related line of work by Alharbi et al. spans two papers: the first retrofits CQs to existing ontologies by prompting LLMs to generate multiple questions per triple [49], and the second extends this with an analysis of model settings and prompt design strategies across LLMs [97]. Both papers are motivated by the needs of ontology engineers, aiming to support ontology testing, reuse, and design evaluation rather than exploratory search. While methodologically thorough, both studies are situated within ontology engineering workflows and do not integrate expert-led curation or address the needs of non-technical users.

There exist recent benchmarks like Bench4KE [47] which focus on automated CQ generation to evaluate ontology quality. This is in contrast to the conceptual framing presented by Tús Maith which (1) generates exploratory scaffold CQ templates as candidate exploratory starting points, and (2) embeds expert curation to ensure usability. This second point is especially crucial in complex, humanities-based graphs like VRTI, where ontology complexity compounds the IEP.

Lastly, recent work on CQ benchmarking [101] provides criteria for evaluating CQ quality across syntactic, semantic, and domain-specific dimensions, including SPARQL transformability. While useful as background for designing evaluation strategies, these frameworks are not yet widely adopted and assume the existence of a GT CQ set, which, up until the work carried out as part of this thesis, did not exist for the VRTI KG.

4.3.4 Stage 2: LLMs and their use in CQ Template Instantiation

LLMs have been applied to KG completion [102] and KG Question-Answering (KGQA) tasks [103], demonstrating their ability to operate over structured semantic information and generate natural language outputs grounded in entity and relation information. These

capabilities made them suitable candidates to investigate as a method of CQ template instantiation for Stage 2 of TMv1. However, prior work also documents issues of hallucination and inconsistency [104]. As a result, as detailed above, the filled-in questions and answers produced during Stage 2 using LLMs are not treated as authoritative outputs but instead, scaffolds for human curation. More recently, complementary work has explored human-in-the-loop architectures to support lay users in KGQA, such as the interactive framework proposed by Pusch et al. [105], which leverages LLMs to generate, explain, and iteratively refine Cypher queries based on natural language feedback. Like Tús Maith, this approach avoids relying on LLMs as authoritative sources by emphasising user oversight and refinement; however, like other KGQA systems, it presupposes a starting point and focuses on query execution rather than CQ template instantiation. Their emphasis on user feedback aligns with TMv1’s scaffolding philosophy, where LLMs facilitate but do not dictate outcomes.

Beyond technical considerations, the design of Stage 2 was informed by human-centred AI and explanatory debugging research [106, 107]. Those works motivated the integration of provenance and light-weight validation design, supporting curator verification rather than automated acceptance. In TMv1, this is presented via the integration of provenance logging, LLM model transparency, a light-weight validation layer, and curator-facing controls that allow domain experts to review and approve entity selections and filled-in questions and answers. This thereby ensures that the LLMs remain as scaffolds in the TMv1 system as opposed to autonomous decision-makers.

4.4 TMv1: System Architecture

TMv1, as an instantiation of Tús Maith with LLMs, at a system architecture level has been designed to be KG-agnostic. However, the concrete instantiation of it as presented in this thesis is using the VRTI KG. The system itself does not assume a specific ontology or domain but instead operates over a set of abstract interface contracts: (1) access to schema-level information in the form of classes, properties, and their admissible relations, (2) an entity ranking function, and (3) a mechanism for materialising the local semantic neighbourhood of a selected entity.

In the implementation described in this thesis, these constraints are fulfilled through SPARQL queries, JSON representations of structured information, and ranking heuristics over the VRTI KG. Accordingly, TMv1, as discussed on a system level, can be understood as an instantiation of the Tús Maith framework using LLMs deployed over the VRTI KG without coupling the core system design to a particular KG.

In the VRTI deployment, TMv1 incorporates an auxiliary mechanism for identifying frequently searched terms, which are used to seed and rank candidate entities for CQ template instantiation. This reflects properties of the VRTI Explorer interface, which records interaction at the level of free-text search terms rather than direct entity selection. A detailed diagram for the system architecture can be seen later in Figure 4.3.

Alternative deployments may realise this function through different means, such as domain expert selection, time-bounded entity sampling, or structural ranking over class-specific

subsets. The system specification below therefore treats popularity-based entity seeding as an optional adapter rather than a core architectural requirement.

4.4.1 External Dependencies

The external dependencies of TMv1 (as deployed using the VRTI KG) are as follows with, when relevant, the expected request and response data format types:

- **VRTI SPARQL Endpoint API:** Utilised in Stage 1 and Stage 2. This API is used to query over the VRTI KG. *Request type:* SPARQL; *Response type:* JSON.
- **Google Analytics API:** Utilised in Stage 2. This is this API from which the analytics information is retrieved from the VRTI Explorer. *Request type:* RunReportRequest; *Response type:* RunReportResponse.
- **LLM API:** Utilised in Stage 1 and Stage 2. This is the API that the prompts specified in the section below are sent to. *Request type:* Plain text prompt. (Expected) *Response type:* JSON.
- **Streamlit [108]:** Utilised as a lightweight software to build the interface with which to hide the technical components of TMv1 in order to present all three stages to the domain experts.

4.4.2 Component Responsibilities

The following components make up TMv1 and are utilised at various points during the three stages of TMv1.

- **KG Adapter Component:** Provides an abstraction layer between TMv1 and the underlying KG. This component is responsible for retrieving schema-level information (classes, properties, and their relationships), validating the existence of entities, and materialising the local semantic neighbourhood of a selected entity. In the VRTI deployment, these functions are realised using SPARQL queries over the VRTI endpoint.
- **Ontology Compaction Component:** Transforms the raw, serialised schema information returned using the KG Adapter component into a reduced and structured representation. This process extracts unique classes, unique properties, and their associated domain and range relationships, producing a compact schema view suitable for use as constrained input to the LLM in Stage 1.
- **LLM Mediation Component:** Serves as the generative core of TMv1. This component is responsible for generating schema-aligned CQ Templates in Stage 1 and for instantiating those templates into filled-in questions and answers in Stage 2, based on the prompts and structured inputs provided by the surrounding components.
- **Stage 1 Prompt Instantiation Component:** Constructs the Stage 1 prompt by combining the compacted schema representation with the fixed prompt template (Appendix A.5.3.1) to guide the LLM in generating CQ templates.

- **Entity Seeding and Ranking Component:** Identifies and ranks candidate entities for use in Stage 2. In the VRTI deployment, this component incorporates popularity signals derived from search activity in the VRTI Explorer, alongside structural heuristics that prioritise lexical overlap between the entity name and the search term, well-connected and information-rich entities.
- **Stage 2 Prompt Instantiation Component:** Constructs the Stage 2 prompt by combining selected CQ templates with materialised entity context and the fixed prompt template (Appendix A.5.3.2) to guide the LLM in generating filled-in questions and answers.
- **Validation Component:** Applies automated checks to filled-in questions and answers, including slot resolution and grounding of answers in the materialised entity context. Outputs are flagged for domain expert review rather than automatically accepted or rejected.
- **CuQA Persistence Component:** Stores CuQAs together with their associated entity identifiers and provenance metadata for deployment within the exploration interface. This is the only component that has not been directly implemented during the course of the research presented as part of this thesis. In spite of having not been implemented, this component is presented here as it would be necessary in full deployment scenarios.

4.4.3 Data and Control Flow

The following canonical example using the VRTI KG represents the end-to-end states of data within TMv1. The presented flow includes internal TMv1 components, human involvement, and external service dependencies. External dependencies are shown where data is retrieved from or sent to systems outside TMv1. Human involvement details the decisions made by domain experts at each stage of TMv1.

Stage 1: CQ Template Derivation

1. SPARQL query → VRTI Endpoint API → JSON (classes, properties, relationships)
2. JSON (classes, properties, relationships) → Ontology Compaction Component → JSON (compact ontology representation)
3. Prompt template + JSON (compact ontology representation) → LLM API → JSON (CQ templates)
4. JSON (CQ templates) → Expert selection → JSON (selected CQ templates)

Stage 2: Template Instantiation

1. RunReportRequest (time period: 7 weeks) → Google Analytics API → RunReportResponse (candidate search terms)

2. SPARQL query → VRTI Endpoint API → JSON (entity existence validation per search term)
3. JSON (candidate seed search terms) → Entity Seeding and Ranking Component → JSON (ranked candidate entities)
4. JSON (ranked candidate entities) → Expert entity selection → JSON (selected entities)
5. SPARQL query → VRTI Endpoint API → JSON (entity semantic neighbourhood materialisation for selected entities)
6. Prompt template + JSON (selected CQ templates) + JSON (selected entities and semantic neighbourhoods) → LLM API → JSON (filled-in questions and answers)
7. JSON (filled-in questions and answers) → Validation Component → Filled-in questions and answers with validation tag

Stage 3: Curation and Persistence

1. Filled-in questions and answers with validation tag → Expert review → CuQA selection
2. CuQA selection → CuQA Persistence Component (entity URI + CuQA + provenance)

One of the operational benefits of Tús Maith utilising an entity centric approach is that maintaining a CuQA collection (post Stage 3, e.g., CuQA RDF Storage in Figure 4.3) requires only an ASK SPARQL to ascertain the continued existence of the entity/entities in question.

4.4.4 Full System Architecture

The full system architecture is presented in Figure 4.3 and is a visual representation of the details of the functional behaviour of TMv1 discussed in the next section.

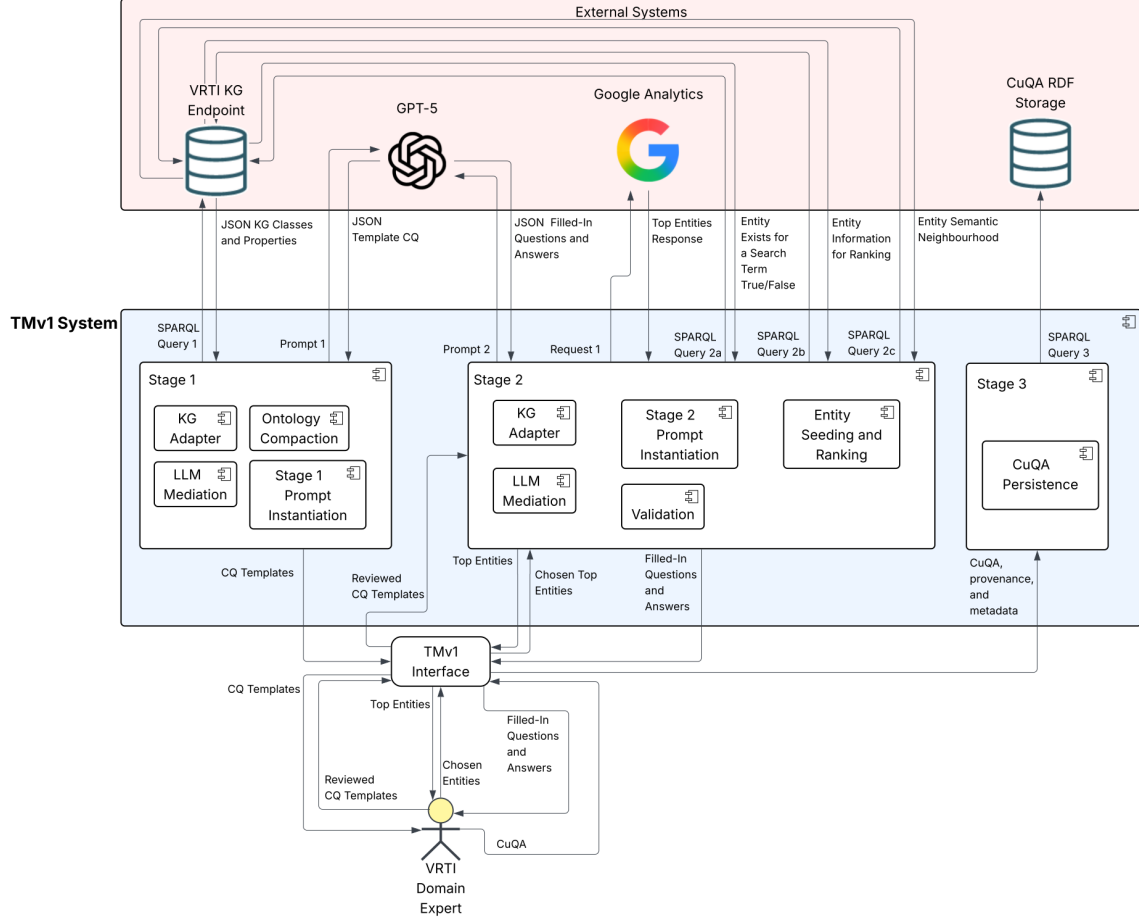


Figure 4.3: TMv1 Component Diagram.

4.5 TMv1: System Specification

This section specifies the functional behaviour of TMv1 at each stage of the Tús Maith framework, detailing the inputs, transformations, and outputs that eventually form CuQAs. The system design of TMv1 presented in this section was informed by User Evaluation 1, User Evaluation 2, LLM Experiment 1, and LLM Experiment 2 detailed in Chapter 5. The results of these evaluations were synthesised to align with the framework’s objectives of supporting non-technical domain experts in curating CuQAs as entry points into a KG and informed the TMv1 prototype implementation to support end-to-end evaluation in User Evaluation 3.

The full code for TMv1, including all reproducibility procedures, can be found in the thesis GitHub repository¹ in the folder User Evaluation 3/T-s_Maith_TMv1. All code was written using Python 3 [109].

4.5.1 Stage 1: LLM-supported CQ Template Derivation

The purpose of Stage 1 of TMv1 is to generate a set of CQ Templates for the given KG. It makes use of the following components: KG Adapter Component, Ontology Compaction Component, and Stage 1 Prompt Instantiation Component.

¹https://github.com/mcnamac1/T-s_Maith_Thesis_Repository

Input: Using the **KG Adapter Component**, a SPARQL query (Appendix A.8.1) is executed over the VRTI SPARQL Endpoint API that retrieves all classes and properties and relationships between them (ontology) of the VRTI KG, returning the ontology in JSON. Listing 4.1 shows an extract of the returned JSON.

Listing 4.1: SPARQL Query Results in JSON from the VRTI KG

```
{
  "head": {
    "link": [],
    "vars": ["subjectClass", "property", "objectClass"]
  },
  "results": {
    "distinct": false,
    "ordered": true,
    "bindings": [
      {
        "property": {
          "type": "uri",
          "value": "http://erlangen-crm.org/current/P12_occurred_in_the_presence_of"
        },
        "objectClass": {
          "type": "uri",
          "value": "http://erlangen-crm.org/current/E21_Person"
        }
      },
      {
        "property": {
          "type": "uri",
          "value": "http://erlangen-crm.org/current/P2_has_type"
        }
      },
      {
        "property": {
          "type": "uri",
          "value": "http://erlangen-crm.org/current/P4_has_time-span"
        },
        "objectClass": {
          "type": "uri",
          "value": "http://erlangen-crm.org/current/E52_Time-Span"
        }
      },
      {
        "property": {
          "type": "uri",
          "value": "http://erlangen-crm.org/current/P89_falls_within"
        },
        "objectClass": {
          "type": "uri",
          "value": "http://www.opengis.net/ont/geosparql#Feature"
        }
      }
    ]
  }
}
```

```

    ...
  ]
}
}

```

As the returned JSON is too verbose to be passed directly to the LLM, it is instead passed into the **Ontology Compaction Component** to compact it, resulting in the output seen in Listing 4.2. More specifically, the **Ontology Compaction Component** processes the raw SPARQL JSON results (as in Listing 4.1) by iterating over each binding. It collects all distinct class URIs that appear either as `subjectClass` or `objectClass` into a sorted list under `unique_classes`. For each property, the component records every unique combination of domain (`subjectClass`) and range (`objectClass`) that occurs in the bindings, storing them as a list of objects under `property_details`. If a binding lacks a subject or object class, that field is set to null in the output. The resulting compact JSON, Listing 4.2, thus preserves all semantic information while eliminating the verbosity of the original SPARQL result format. The full output for both Listing 4.1 and Listing 4.2 can be found in this thesis' GitHub repository ² in the folder User Evaluation 3/T-s_Maith_TMv1.

Listing 4.2: Compact Ontology JSON (Partial)

```

{
  "unique_classes": [
    "http://erlangen-crm.org/current/E21_Person",
    "http://erlangen-crm.org/current/E24_Physical_Human-Made_Thing",
    "... 11 more classes ...",
    "http://erlangen-crm.org/current/E67_Birth",
    "http://erlangen-crm.org/current/E69_Death",
    "http://erlangen-crm.org/current/E71_Human-Made_Thing",
    "http://erlangen-crm.org/current/E81_Actor_Appellation*",
    "http://www.opengis.net/ont/geosparql#Feature",
    "http://www.opengis.net/ont/geosparql#Geometry"
  ],
  "unique_properties": [
    "http://data.logainm.ie/ontology/easting",
    "http://data.logainm.ie/ontology/northing",
    "http://data.logainm.ie/ontology/square",
    "http://erlangen-crm.org/current/P102_has_title",
    "http://erlangen-crm.org/current/P107_has_current_or_former_member",
    "http://erlangen-crm.org/current/P12_occurred_in_the_presence_of",
    "... 10 more properties ...",
    "https://www.w3id.org/virtual-treasury/ontology#century",
    "https://www.w3id.org/virtual-treasury/ontology#has_era_type",
    "https://www.w3id.org/virtual-treasury/ontology#links"
  ],
  "property_details": {
    "http://erlangen-crm.org/current/P12_occurred_in_the_presence_of": [
      {
        "domain": null,

```

²https://github.com/mcnamacl/T-s_Maith_Thesis_Repository

```

        "range": "http://erlangen-crm.org/current/E21_Person"
    },
],
"http://erlangen-crm.org/current/P2_has_type": [
    {
        "domain": null,
        "range": null
    },
    "... 7 more domain-range pairs ...",
    {
        "domain": "http://www.opengis.net/ont/geosparql#Feature",
        "range": null
    }
],
"http://erlangen-crm.org/current/P4_has_time-span": [
    {
        "domain": null,
        "range": "http://erlangen-crm.org/current/E52_Time-Span"
    },
    {
        "domain": "http://erlangen-crm.org/current/E67_Birth",
        "range": "http://erlangen-crm.org/current/E52_Time-Span"
    },
    {
        "domain": "http://erlangen-crm.org/current/E69_Death",
        "range": "http://erlangen-crm.org/current/E52_Time-Span"
    }
],
"... additional property mappings ..."
}
}

```

**Should be E82_Actor_Appellation, however, at the time it was entered as E81 in the VRTI KG.*

Transformation: The **Stage 1 Prompt Instantiation Component** takes this compact ontology and inserts it within the Stage 1 Prompt (Appendix A.5.3.1). The **LLM Mediation Component** feeds the prompt to the LLM, in this instance Open AI's GPT-5 [53], and asks it to generate 20 CQ Templates. It then interprets the JSON response extracting the 20 CQ Templates.

Output: The resulting output is 20 CQ Templates that have been generated via the LLM using the compact ontology of the KG. A canonical example of the output as used as part of User Evaluation 3 can be seen in Listing 4.3.

Listing 4.3: Template Competency Questions for the VRTI KG

```

{
  "questions": [
    "Where was a person born?",
    "When was a person born?",

```

```

    "Where did a person die?",
    "When did a person die?",
    "What are the earliest and latest possible dates for a person's birth?"
    ,
    "Where else on the web can I find a person?",
    "What names does a person go by?",
    "What alternative versions of a person's name are recorded?",
    "Which images or objects depict a person?",
    "Which sources list a person?",
    "Which people are listed in an authority source?",
    "Which sources refer to a place?",
    "Who contributed to an authority source?",
    "What century is a person associated with?",
    "What era is a place associated with?",
    "What is the preferred name of a place?",
    "What official identifiers does a place have?",
    "What larger area does a place fall within?",
    "What is the area, centroid, or bounding box of a place?",
    "What is the ISBN or web link for an authority source?"
  ]
}

```

4.5.2 Stage 2: LLM-supported CQ Template Instantiation

Stage 2's purpose is to instantiate the CQ Templates generated from Stage 1 with specific entities from the KG. It utilises the following components: KG Adapter Component, Entity Seeding and Ranking Component, LLM Mediation Component, Stage 2 Prompt Instantiation Component, and the Validation Component.

Input: Stage 2 takes in the CQ templates that have been selected from the set generated as part of Stage 1. It furthermore uses the **Entity Seeding and Ranking Component** to interface with the Google Analytics API to retrieve the top search terms over a specified time period from the KG Exploration Interface (in this case, the VRTI Explorer). Using those search terms and the **KG Adapter Component** with an ASK SPARQL query (Appendix A.8.2), the **Entity Seeding and Ranking Component** heuristically ranks entities whose names match with the search term according to how well the name matches the search term (lexical overlap after normalisation), how richly connected the entity is (in- and out-degree), and how much information is available for the entity (number of slots of information present) using the SPARQL queries presented in Appendix A.8.3. The semantic neighbourhood for the top ranked entities is extracted using the **KG Adapter Component** with a SPARQL query (Appendix A.8.4) ready for selection by the domain expert.

Transformation: The **Stage 2 Prompt Instantiation Component** inserts selected CQ templates and selected entities into the Stage 2 prompt (Appendix A.5.3.2). The **LLM Mediation Component** then sends the prompt to the LLM (once again, GPT-5 [53]), asking it to instantiate the CQ templates from Stage 1 with the entities chosen earlier as

part of of Stage 2. An example of the entity information passed to the LLM can be seen in Listing 4.4.

Listing 4.4: Entities with their information from VRTI KG

```
{
  "https://kg.virtualtreasury.ie/person/Ward_Maxwell-Richard-Crosbie_c20/v1mp8q3": [
    {
      "Person": "https://kg.virtualtreasury.ie/person/Ward_Maxwell-Richard-Crosbie_c20/v1mp8q3",
      "Gender": "https://www.w3id.org/virtual-treasury/ontology#Male",
      "Surname": "https://kg.virtualtreasury.ie/normalized-appellation-surname/Ward",
      "FullNameLabel": "Ward, Maxwell Richard Crosbie",
      "FirstNameLabel": "Maxwell Richard Crosbie",
      "DIB_Page": "https://www.dib.ie/biography/Ward-Maxwell-Richard-Crosbie-a8922",
      "TimePeriod": "https://www.w3id.org/virtual-treasury/vocabulary#Modern-Post-1922",
      "DeathDateUpper": "1950-11-17",
      "DeathDateLower": "1950-11-17",
      "DisplayDeathDate": "17-11-1950",
      "BirthDateUpper": "1868-05-04",
      "BirthDateLower": "1868-05-04",
      "DisplayBirthDate": "4-5-1868",
      "DisplayFloruitLower": "4-5-1868",
      "DisplayFloruitUpper": "17-11-1950",
      "FloruitLowerYear": "1868",
      "FloruitUpperYear": "1950",
      "SurnameLabels": "Ward",
      "DeathPlace": "http://www.wikidata.org/entity/Q26;Northern Ireland---https://kg.virtualtreasury.ie/place/present-day/county/Down/v15h6sz;Down",
      "BirthPlace": "",
      "Image": "http://commons.wikimedia.org/wiki/Special:FilePath/Maxwell%20Richard%20Crosbie%20Ward%2C%206th%20Viscount%20Bangor.jpg"
    }
  ],
  "https://kg.virtualtreasury.ie/place/present-day/townland/Wardstown/v1mj2c9": [
    {
      "CRM_Type": "https://ont.virtualtreasury.ie/ontology#PresentDayTownland",
      "LogainmLink": "http://data.logainm.ie/place/39165",
      "TownlandsIELink": "http://www.townlands.ie/meath/lune/athboy/rathmore/tlachta",
      "Centroid": "https://kg.virtualtreasury.ie/place/present-day/townland/centroid/Wardstown/v1mj2c9",
      "Point": "POINT(53.6269365558 -6.8772113537)",
      "Names": "Tlachta---Wardstown",
      "FallsWithinPlaces": "https://kg.virtualtreasury.ie/place/present-day/parish/Athboy/v1w9hp3",
    }
  ]
}
```

```

    "FallsWithinPlaceTypes": "https://ont.virtualtreasury.ie/ontology#
      PresentDayParish",
    "FallsWithins": "https://kg.virtualtreasury.ie/place/present-day/
      parish/Athboy/v1w9hp3;Athboy",
    "FallsWithinPlaceLabels": "Athboy",
    "parentLabels": "An Mh ---Athboy---Baile tha Bu ---C ige
      Laighean---Leinster---Lune---Lu ne ---Meath"
  }
],
"https://kg.virtualtreasury.ie/place/present-day/townland/Ward-Upper/
  v166snz": [
  {
    "CRM_Type": "https://ont.virtualtreasury.ie/ontology#
      PresentDayTownland",
    "LogainmLink": "http://data.logainm.ie/place/17332",
    "TownlandsIELink": "http://www.townlands.ie/dublin/castleknock/ward/
      finglas-ed-1901/ward-upper",
    "Centroid": "https://kg.virtualtreasury.ie/place/present-day/townland
      /centroid/Ward-Upper/v166snz",
    "Point": "POINT(53.4392663249 -6.3447457869)",
    "Names": "Ward Upper",
    "FallsWithinPlaces": "https://kg.virtualtreasury.ie/place/present-day
      /parish/Ward/v1r7np8",
    "FallsWithinPlaceTypes": "https://ont.virtualtreasury.ie/ontology#
      PresentDayParish",
    "FallsWithins": "https://kg.virtualtreasury.ie/place/present-day/
      parish/Ward/v1r7np8;Ward",
    "FallsWithinPlaceLabels": "Ward",
    "parentLabels": "An Barda---Baile tha Cliath---Caisle n Cnucha---
      Castleknock---C ige Laighean---Dublin---Leinster---Ward"
  }
]
}

```

The returned output from the LLM should be a JSON object comprising of a set of filled-in questions and answers with, for each question and answer, the URI of the entity (provenance) used to instantiate the CQ template to produce the filled-in question and answer. Each filled-in question and answer is passed through the validation component to ensure that there is provenance attached, that the entity name and the name used to fill in the CQ template match after normalisation, and that the answer is information contained within the semantic neighbourhood of the entity.

Output: The final output is a set of filled-in questions and answers that have been instantiated using an LLM. An example of the final output using a person and a place URI can be seen in Listing 4.5 below.

Listing 4.5: Filled-In Questions and Answers

```

{
  "results": [
    {

```

```

    "question": "Which sources list Ward, Maxwell Richard Crosbie?",
    "answer": "Dictionary of Irish Biography; James McGuire and James
               Quinn (eds), Dictionary of Irish Biography (Cambridge, 2009);
               https://www.dib.ie/biography/Ward-Maxwell-Richard-Crosbie-a8922",
    "provenance": {
      "source_entity_uri": "https://kg.virtualtreasury.ie/person/
                           Ward_Maxwell-Richard-Crosbie_c20/v1mp8q3"
    }
  },
  {
    "question": "What larger area does Ward Upper fall within?",
    "answer": "Ward (An Barda)",
    "provenance": {
      "source_entity_uri": "https://kg.virtualtreasury.ie/place/present-
                           day/townland/Ward-Upper/v166snz"
    }
  },
  {
    "question": "What larger area does Wardstown fall within?",
    "answer": "Athboy (Baile tha Bu )",
    "provenance": {
      "source_entity_uri": "https://kg.virtualtreasury.ie/place/present-
                           day/townland/Wardstown/v1mj2c9"
    }
  },
  {
    "question": "Which sources refer to Ward Upper?",
    "answer": "http://www.townlands.ie/dublin/castleknock/ward/finglas-ed
               -1901/ward-upper; https://www.logainm.ie/en/17332",
    "provenance": {
      "source_entity_uri": "https://kg.virtualtreasury.ie/place/present-
                           day/townland/Ward-Upper/v166snz"
    }
  },
  {
    "question": "Which sources refer to Wardstown?",
    "answer": "https://www.logainm.ie/en/39165; http://www.townlands.ie/
               meath/lune/athboy/rathmore/tlachta; https://fionstorage.blob.core
               .windows.net/topoimages/Box-3939165_1.JPG; https://fionstorage.
               blob.core.windows.net/topoimages/Box-3939165_2.JPG; https://
               fionstorage.blob.core.windows.net/topoimages/Box-3939165_3.JPG;
               https://fionstorage.blob.core.windows.net/topoimages/Ref-2939165.
               jpg",
    "provenance": {
      "source_entity_uri": "https://kg.virtualtreasury.ie/place/present-
                           day/townland/Wardstown/v1mj2c9"
    }
  }
]
}

```

4.5.3 Stage 3: Expert Curation

The final stage is where the domain experts have final control over whether or not a filled-in question and answer meets the criteria to become a Curated Question and Answer (CuQA). Each of the filled-in questions and answers are inspected by them via a user interface, further detail of which, and figures showing what the interface looked like, are presented in User Evaluation 3, in Chapter 5, Section 5.9. They can choose if the question is relevant, answerable, and clear and therefore is a candidate to be an exploratory entry point into the KG. This final stage makes use of the following component: CuQA Persistence Component.

Input: The input is the filled-in questions and answers created by Stage 2.

Transformation: If a filled-in question and answer is deemed suitable to be published as an exploratory starting point for lay users, it is promoted to *CuQA*.

Output: The output is a set of CuQAs that expose the scope of the KG to lay users and bring those lay users to richly connected areas of the KG.

The **CuQA Persistence Component** stores the CuQA using RDF with URI/s used to instantiate the CuQA, provenance, and associated metadata about the CuQA (datetime of creation, curator, etc.). While, as previously mentioned, this component was not implemented as part of TMv1 due to lack of necessity (none of the evaluations required persistent storage), an example of how the information could be stored is presented here. Such provenance information can be stored using the PROV-O [110] ontology which provides classes and properties that describe the CuQA (`prov:Entity`), CQ template (`prov:Entity`), curator who carried out the act of curation (`prov:Agent`), and the act of curation itself (`prov:Activity`). The relationship between the curator and the CuQA can be modelled using `prov:wasAttributedTo`, the link between the curation activity and the CuQA using `prov:wasGeneratedBy`, the datetime of curation recorded using `prov:generatedAtTime`, and source entity (entities) and CQ template modelled both using `prov:wasDerivedFrom`. An illustrative example of what this would look like storing the CuQA as RDF using PROV-O can be seen in Listing 4.5.1.

Listing 4.5.1: CuQA Persistence Component: Storing CuQAs using PROV-O.

true true

```
1 @prefix prov: <http://www.w3.org/ns/prov#> .
2 @prefix tm:   <http://example.org/tusmaith/> .
3 @prefix foaf: <http://xmlns.com/foaf/0.1/> .
4 @prefix xsd:  <http://www.w3.org/2001/XMLSchema#> .
5 @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
6
7 ### CQ TEMPLATE
8 tm:cqTemplate_1 a prov:Entity ;
9     rdfs:label "What larger area does a place fall within?" .
10
11 ### CURATED QUESTION AND ANSWER (CuQA)
```

```

12 tm:cuqa_1 a prov:Entity ;
13     rdfs:label "What larger area Ward Upper fall within?" ;
14     tm:answer "Ward (An Barda)" ;
15     prov:wasDerivedFrom tm:cqTemplate_1 ;
16     prov:wasDerivedFrom
17         ↪ <https://kg.virtualtreasury.ie/place/present-day/townland/Ward-Upper/v166snz> ;
18     prov:wasAttributedTo tm:curator_1 ;
19     prov:wasGeneratedBy tm:curation_activity_1 ;
20     prov:generatedAtTime "2026-02-09T14:22:00Z"^^xsd:dateTime .
21
22 ### CURATION ACTIVITY
23 tm:curation_activity_1 a prov:Activity ;
24     prov:wasAssociatedWith tm:curator_1 ;
25     prov:used tm:cqTemplate_1 ;
26     prov:used
27         ↪ <https://kg.virtualtreasury.ie/place/present-day/townland/Ward-Upper/v166snz> ;
28     prov:endedAtTime "2026-02-09T14:22:00Z"^^xsd:dateTime .
29
30 ### CURATOR
31 tm:curator_1 a prov:Agent ;
32     foaf:name "Domain Expert 1" ;
33     tm:curatorID "DE1" .

```

4.6 TMv1 Prototype Configuration for User Evaluation 3

As noted in Section 4.4.1, TMv1 was exposed to domain experts via a lightweight prototype interface implemented using Streamlit [108]. This interface was used for User Evaluation 3, which assessed the end-to-end use of TMv1 by domain experts. Full details of the evaluation design and results are presented in Chapter 5.

This section specifies the exact operational configuration of TMv1 during User Evaluation 3. In order to focus the evaluation on the three stages of the Tús Maith framework as instantiated by TMv1, certain components of the system were precomputed (frozen) or computed after the evaluation, while others were executed live during expert interaction. This distinction is important: while the full implemented TMv1 system includes all components shown in Figure 4.4, domain experts interacted only with the live subset of these components via the prototype interface as seen in Figure 4.5. The configuration described here reflects empirically informed design decisions discussed in Chapter 5.

The evaluated configuration of TMv1 was as follows:

- **LLM:** GPT-5 [53], used for both precomputed and live generation with parameters set to `temperature=1` and `top_p=1`.
- **Stage 1 (CQ Template Derivation):** Stage 1, as shown in Figure 4.4, was precomputed prior to the evaluation. A set of 20 CQ templates was generated using GPT-5 from the compact ontology derived from the classes, properties, and relationships of the VRTI KG as retrieved on 20 November 2025, reflecting the state of the KG at that time. From this set, six CQ templates (three relating to people and

three relating to places) were selected by the author and presented to domain experts during the evaluation.

- **Stage 2 (Entity Seeding):** Candidate entities were precomputed prior to the evaluation using the Entity Seeding and Ranking component. The top five search terms recorded in the VRTI Explorer search box during the 21 weeks following public launch (30 June 2025 to 26 November 2025) were used: $\{alex, james, ward, maguire, kelly\}$. For each search term, the top ten associated entities (five people and five places) were generated in advance and made available for expert selection within the prototype.
- **Stage 3 (CQ Template Instantiation):** CQ templates were instantiated live by GPT-5 based on the templates and entities selected by the domain expert during the evaluation.

To provide a complete architectural overview, Figure 4.4 shows the full TMv1 system as implemented, including frozen (precomputed) artefacts, live components, and the component used post-evaluation (Validation Component). This represents the complete operational system underpinning the evaluation.

In contrast, Figure 4.5 shows only the subset of TMv1 components that were executed live and directly experienced by domain experts through the prototype interface during User Evaluation 3. It follows the same layout as the previous figure (Figure 4.4), but has removed the components not utilised in the live version and instead, presents the artefacts precomputed using the full system in yellow.

4.7 Summary

This chapter has presented the design of the Tús Maith framework and its instantiation and implementation as TMv1 deployed using the VRTI KG. In doing so, it has addressed RQ2 by demonstrating how the conceptual lifecycle of CuQA generation can be implemented as a working system that supports non-technical domain experts in creating exploratory entry points into a KG.

The chapter first established the conceptual foundations of Tús Maith as a three-stage lifecycle for generating, instantiating, and curating natural-language questions and answers (CuQAs). It articulated the epistemic and interactional rationale for treating curated questions as exploration primitives and introduced an entity-centric anchoring model that distinguishes exploratory interaction from traditional query translation approaches.

It then described the instantiation of this framework as TMv1, justifying the use of LLM-supported generation and detailing the oversight, validation, and provenance mechanisms required to ensure reliability. The system architecture and component responsibilities were specified in sufficient detail accompanied by a comprehensive GitHub repository to support reproducibility and to demonstrate how the framework can be realised while remaining agnostic to domain and KG.

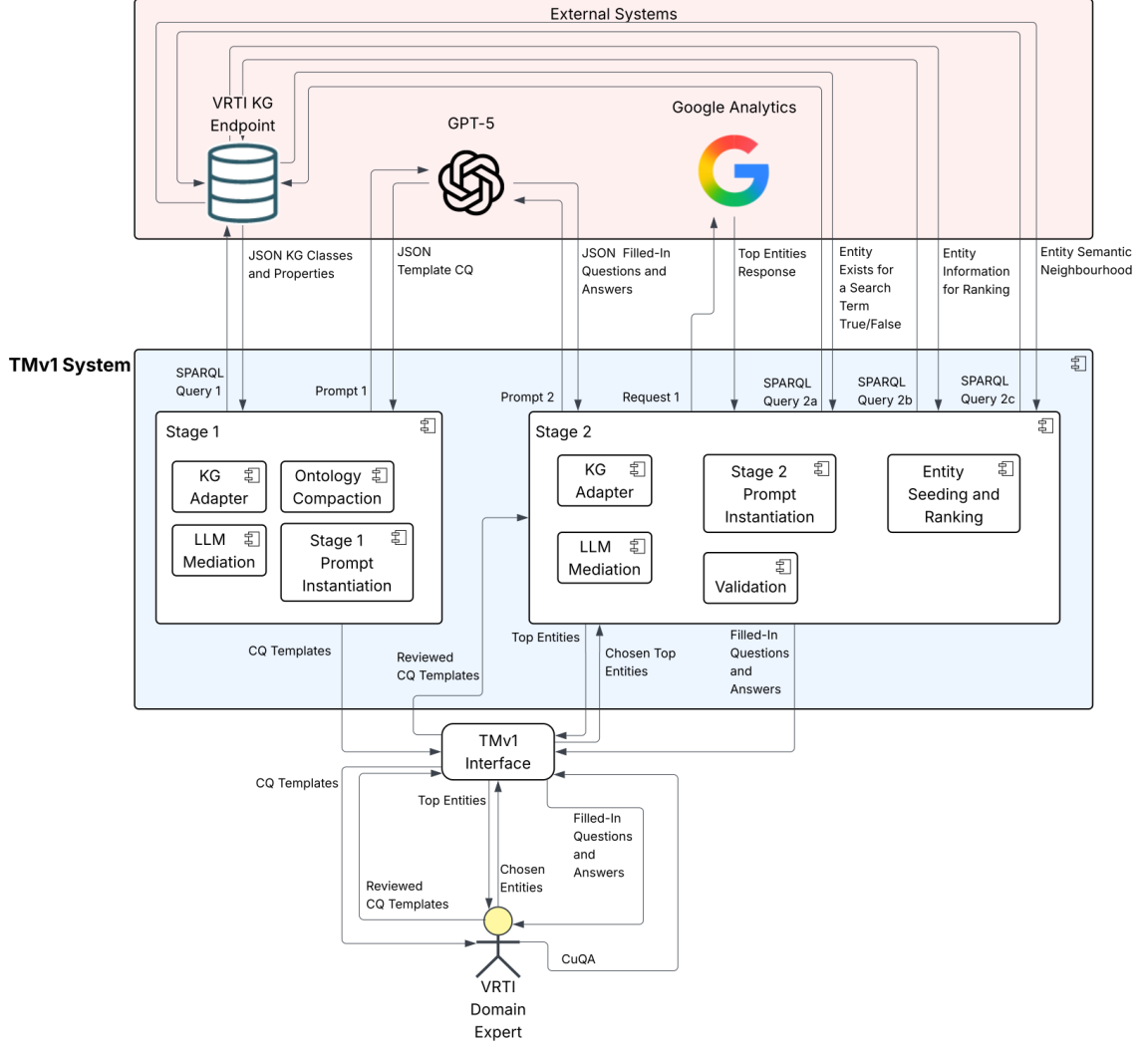


Figure 4.4: Full TMv1 system configuration showing both frozen (precomputed) and live components for User Evaluation 3.

Finally, the chapter documented the evaluated configuration of TMv1 used in User Evaluation 3, establishing a clear correspondence between conceptual design, technical implementation, and empirical assessment.

Having established both the conceptual and technical foundations of Tús Maith and its instantiation as TMv1, the following Evaluation Chapter, Chapter 5, examines the extent to which CuQAs function as exploratory entry points (User Evaluation 1) and assesses the degree to which CuQA mitigate the IEP for lay users (User Evaluation 2), how LLMs function in Stage 1 and Stage 2 of TMv1 (LLM Experiment 1 and LLM Experiment 2), how viable TMv1, as deployed using the VRTI KG, is considered by qualified domain experts (User Evaluation 3), and finally, a proof-of-concept of Tús Maith, as instantiated as TMv1, using a different KG.

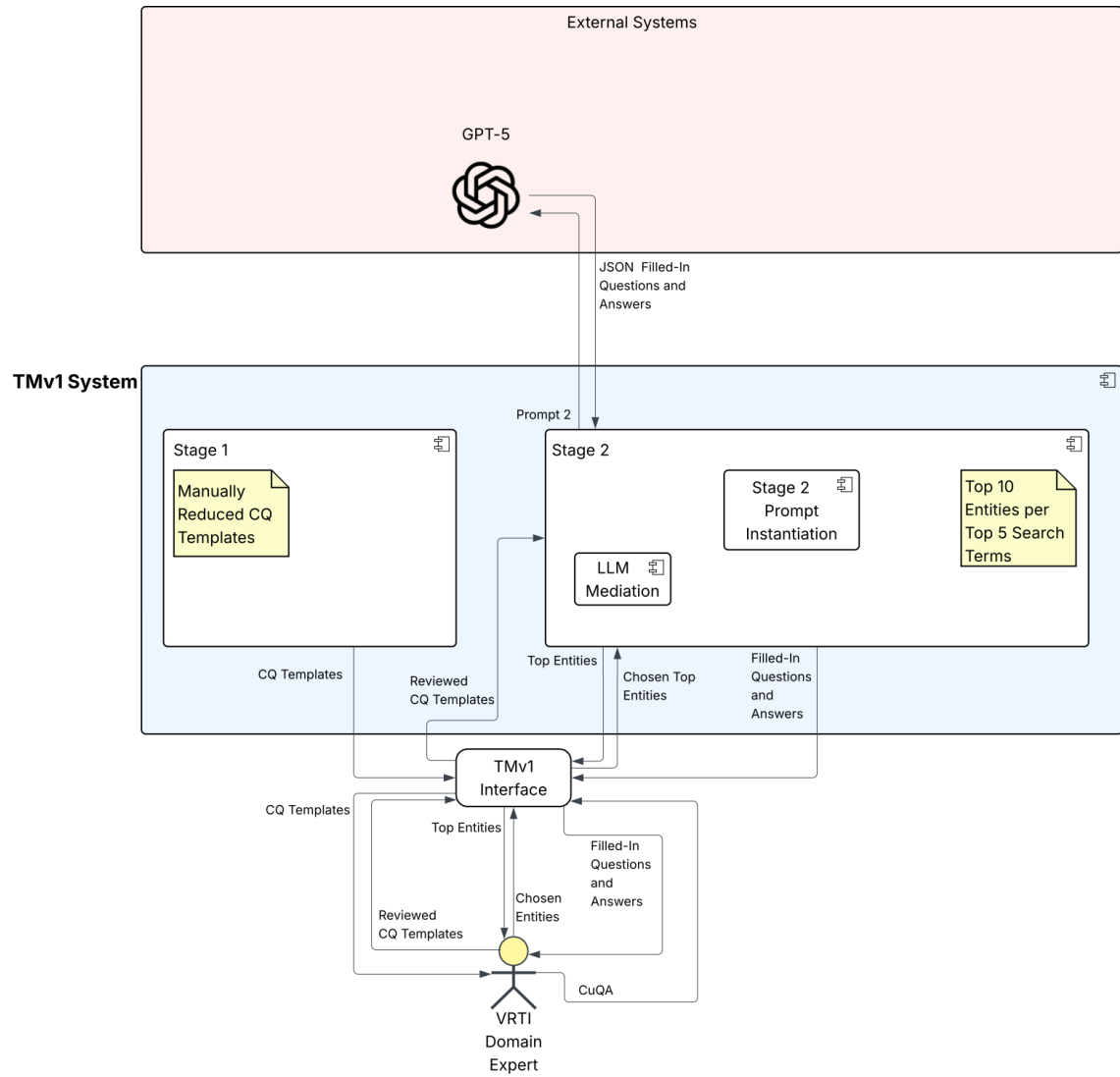


Figure 4.5: Live components of TMv1 as experienced by domain experts during User Evaluation 3.

Chapter 5

Evaluation

The Evaluation chapter starts by providing an overview of the five evaluations undertaken as part of this thesis. It then presents the results of the evaluations and explores the extent to which analysis of the results provides evidence in support of addressing the research question under examination. Section 5.1 describes the purpose of the evaluations, Section 5.2 details the various instruments and analysis methods used to obtain and interpret the findings of the five evaluations, and Section 5.3 provides context necessary to understand how the evaluations were conducted. Section 5.4 presents the how the overlap in domain experts as curators and participants was handled. Sections 5.5 to 5.9 describe in detail the design, implementation, and results of each of the five evaluations. Section 5.10 presents a proof-of-concept of TMv1 applied to a separate digital humanities KG. Finally, Section 5.11 synthesises and discusses the insights gained through the evaluations and offers overall conclusions drawn from the results.

5.1 Evaluation Overview and Rationale

5.1.1 Evaluation Goals

The evaluation goals here are three-fold. Firstly, to assess whether CuQAs can function as entry-points into an unfamiliar, large, complex KG and, the extent to which they address the Initial Exploration Problem (IEP), (RO3a). Where a complex KG here is understood as: one whose underlying ontology introduces semantic and structural complexity that makes the graph difficult to interpret and explore without prior ontological or technical expertise. Secondly, to assess the extent to which LLMs can carry out Stage 1 (CQ Template Derivation) and Stage 2 (CQ Template Instantiation) of TMv1 (RO3b). Finally, to assess the extent to which a human-in-the-loop LLM-supported system, TMv1, can support domain experts in producing CuQAs (RO3c).

5.1.2 Mixed-Methods Evaluation Strategy

In order to holistically examine solutions put forward to address the research question of the thesis, a mixed-methods evaluation strategy was adopted. This strategy encompasses two complementary strands of evaluation: user studies and technical experiments. The approach as a whole can be considered a pragmatic, iterative one with each evaluation building on the lessons learned from the previous.

User Studies The characteristics of the IEP are not only observable outcomes, but also incorporate users’ cognitive processes, interpretations, and perceptions when engaging with unfamiliar knowledge structures. As such, reliance on a single methodological lens would provide only a partial account. Quantitative measures alone are insufficient to surface the reasoning and sense-making processes that underpin user behaviour, while qualitative data in isolation cannot adequately capture the extent or consistency of perceived effects across participants [111].

Accordingly, this thesis employed a convergent mixed-methods design [112], in which quantitative and qualitative data were collected within the same evaluation and analysed in parallel. This approach enabled complementary perspectives on user experience and perceived support, while allowing different forms of evidence to inform and contextualise one another.

Technical Experiments Explanatory sequential design [112] was adopted for LLM Experiment 1 in order to narrow output through automated measures and obtain targeted human feedback.

5.1.3 Mapping of Evaluations to Research Objectives

Table 5.1 presents how each of the five evaluations maps to the research objectives detailed in Chapter 1.

Table 5.1: Overview of evaluations and their corresponding research objectives

Evaluation	Research Objective	Evaluation Aim
User Evaluation 1 (UE1)	RO3a	To understand the extent to which CuQAs can act as entry points into an unfamiliar KG.
User Evaluation 2 (UE2)	RO3a	To understand the extent to which CuQAs address the IEP.
LLM Experiment 1 (EX1)	RO3b	To understand the extent to which LLMs can produce CQ templates suitable for Tús Maith.
LLM Experiment 2 (EX2)	RO3b	To understand the extent to which LLMs can effectively perform batch instantiation of CQ templates in the generation of questions and answers across multiple entities while maintaining accuracy, fidelity, and robustness.
User Evaluation 3 (UE3)	RO3c	To understand the extent to which TMv1 aids domain experts in the efficient creation of publishable CuQAs for the purpose of Tús Maith.

This sequence reflects a development logic: foundational investigations into CuQA efficacy (UE1, UE2) and LLM capability (EX1, EX2) directly informed the design and configuration of the final system (TMv1 prototype), whose viability was then assessed with domain experts (UE3). Throughout the evaluations, in order to differentiate the evaluation research questions from the thesis’ research question, the LLM experiments have been given the acronym RQI_EXJ and the user evaluation evaluations have been given the acronym RQI_UEJ, where I denotes the question’s number, and J denotes the evaluation/experiment number.

5.2 Evaluation Instruments and Analysis Methods

This section presents the various quantitative and qualitative methods used to capture and analyse the results of all five evaluations.

5.2.1 User Evaluation Instruments

5.2.1.1 Questionnaires

Standardised and custom questionnaire instruments were utilised to capture a variety of constructs including workload, system trust, and artefact-specific dimensions of user experience when users were interacting with CuQAs (User Evaluation 1 and User Evaluation 2) and TMv1 evaluation (User Evaluation 3).

NASA Raw Task Load Index (RTLX). The NASA Raw Task Load Index (RTLX) [113] was used to assess perceived cognitive and operational workload across five dimensions: mental demand, temporal demand, performance, effort, and frustration. Following [114], RTLX was modified to remove the question relating to physical demand as it was not deemed relevant to the evaluations undertaken as part of this thesis. Participants rated each dimension on a discrete scale 1 - 5 following task completion. These measures provided a structured, quantitative account of the cognitive burden associated with utilising TMv1. RTLX scores were analysed both at the subscale level and as an aggregated workload construct to enable comparisons across evaluation constructs.

Short Trust in Intelligent Automation Scale (S-TIAS). The Short Trust in Intelligent Automation Scale (S-TIAS) [115] was used to measure user trust in TMv1 as an AI-assisted system. The instrument captures perceptions of system reliability, predictability, and dependability, which are particularly relevant in curator-facing systems where users must decide whether to accept, revise, or reject system-generated content. S-TIAS responses were analysed to assess how provenance presentation and validation feedback influenced trust in an AI-assisted system.

Custom Questionnaire Items. In addition to standardised instruments, a set of custom-designed questions was developed to capture evaluation dimensions not addressed by existing scales. These items focused on three primary areas:

- **CuQA Assessment:** Items probing perceived relevance, clarity, and answerability of CuQA artefacts as entry points into the VRTI KG (more detail on this is provided in LLM Experiment 1, Section 5.7).
- **Provenance and Trust:** An item examining how the presence of provenance information (e.g., source URIs) affected participants’ trust in the correctness and interpretability of generated CuQAs.
- **Interface Feature Ranking:** Ranking and Likert-style items used to elicit user priorities among interface features.

Custom items were analysed using descriptive statistics and, where applicable, inferential tests to identify systematic differences across evaluation conditions and participant groups.

5.2.1.2 Think-Aloud Protocol

A guided, moderated concurrent verbalisation protocol [116] was employed in User Evaluation 3 to capture participants’ reasoning and decision-making during interaction with TMv1 and to contextualise observed behavioural and questionnaire outcomes. The purpose of this protocol was explanatory rather than exploratory. In other words, it was used to help interpret how and why participants interacted with TMv1 in the ways reflected in behavioural metrics and post-task responses rather than to capture fully independent or unprompted cognitive strategies.

Unlike a purely self-directed think-aloud procedure, participants were not expected to complete the task independently. Instead, the author of this thesis followed a predefined task script and provided step-by-step explanations of what was expected at each stage of the TMv1 system, as well as clarifications in response to participant questions. This structured approach ensure procedural consistency across sessions. Participants were encouraged to articulate their thoughts as they progressed through each scripted stage, including their use of provenance information, and their rationale for editing candidate CuQAs. The moderator’s role was primarily procedural, in ensuring consistent task progression across participants but also probative in a limited sense, prompting clarification where necessary to understand participants actions (for example, questioning participants to understand why they clicked to view provenance).

All sessions were audio-recorded and transcribed using a combination of automated transcription [117] and manual correction. Given the interactive and instructional nature of the protocol, the resulting transcripts were not treated as an independent source of unprompted cognitive strategies or subjected to formal inductive or thematic coding. Instead, they were used as an explanatory resource to contextualise behavioural metrics and questionnaire responses, and to surface illustrative examples of user reasoning, points of confusion, and trust calibration behaviour observed during system use.

5.2.1.3 Behavioural Metrics

In addition to self-reported measures during User Evaluation 3, system-level interaction logs were collected to provide objective behavioural indicators of user engagement and

curation effort. The following metrics were captured:

- **Time per TMv1 Stage:** The duration spent in each stage of TMv1, used as a proxy for interactional effort and efficiency.
- **Provenance Interactions:** The number of clicks to view provenance interpreted as an indicator of verification behaviour and trust calibration.
- **CuQA Edits:** The frequency of edits made to generated CuQAs.
- **Publishable CuQAs:** The number of filled-in questions and answers deemed acceptable for publication (transforming them into CuQAs) by domain experts, used as an outcome-oriented measure of system utility and output quality.
- **Percentage of CuQAs:** The number of CuQAs / total number of generated filled-in questions and answers used as an outcome-oriented measure of system utility and output quality.

These behavioural measures were analysed alongside questionnaire and qualitative data for User Evaluation 3 to triangulate perceived system performance with observed user behaviour.

5.2.1.4 Qualitative Analysis Methods

Qualitative material in User Evaluation 3 consisted of transcripts from the guided, moderated verbalisation protocol. All three user evaluations made use of open-ended questionnaire responses. Neither the data obtained from the transcripts, nor the questionnaire comments, were treated as an independent corpus for inductive or thematic analysis. Instead, they were examined using a directed, explanatory review approach, in which excerpts were selected to contextualise and interpret quantitative and behavioural findings rather than to generate standalone qualitative claims.

For User Evaluation 3, transcript analysis proceeded by identifying segments of participant commentary that corresponded to predefined analytical dimensions derived from the evaluation goals, including perceived workload, trust calibration, use of provenance, interpretation of validation feedback, and criteria for accepting or rejecting candidate CuQAs. These segments were used to construct illustrative examples and explanatory narratives that help clarify why particular behavioural patterns or questionnaire trends were observed. To maintain transparency and reduce selective reporting, excerpts were drawn across participants and across task stages, with attention to both convergent and divergent perspectives. The resulting qualitative evidence is therefore presented as supporting and interpretive rather than as formally coded or generalisable findings.

5.2.2 LLM Evaluation Instruments

LLM performance was assessed in terms of the quantitative and light qualitative (expert open-ended comments) instruments used to measure how effectively LLM models operated

within Stage 1 and Stage 2 of TMv1. Rather than treating LLMs as generic CQ template derivators and instantiators, the metrics were defined to capture their suitability as *scaffolding components* within a curator-facing CuQA generation system.

Evaluation instruments were organised according to the two functional roles of LLMs in TMv1: (1) CQ Template Derivation (LLM Experiment 1, Stage 1, Section 5.7), and (2) CQ Template Instantiation to produce filled-in questions and answers using in KG entities (LLM Experiment 2, Stage 2, Section 5.8).

Stage 1: CQ Template Derivation Metrics Performance in Stage 1 was measured using a combination of semantic alignment metrics and domain expert judgement designed to capture both coverage (the extent to which the CQ templates describe the scope of the KG) and qualitative suitability of generated CQ templates.

- **Semantic Similarity (SentenceBERT (SBERT) Cosine Similarity [42]):** For each generated CQ template, the maximum cosine similarity to any expert-authored ground truth (GT) CQ was computed. Aggregate measures included mean maximum similarity per run and the proportion of questions exceeding a fixed similarity threshold ($\tau = 0.6$, chosen following prior work from [99]). This metric utilises semantic anchoring to expert-defined KG coverage.
- **Relevance (Likert, 1-5):** Domain expert rating of how well a generated CQ template aligns with the intended purpose of CQ templates being repurposed (once instantiated) as an exploratory entry point into the KG.
- **Clarity (Likert, 1-5):** Domain expert rating of linguistic quality, including grammaticality, ambiguity, and interpretability for non-technical users.
- **Answerability (Likert, 1-5):** Domain expert rating of whether a generated CQ template can reasonably be answered using only the information present in the VRTI KG.
- **Coverage Expansion (Binary, Expert Judgment):** Whether a generated CQ template was judged by an expert to be suitable for inclusion in the ground truth (GT) set of CQ templates (a set of CQ templates created by a domain expert which they considered captured the breadth of the KG), capturing the model’s ability to surface plausible questions beyond domain expert-authored coverage.
- **Inter-Rater Reliability (ICC(3,k)) [118]:** Agreement between domain experts across relevance, clarity, and answerability ratings, used to quantify the consistency and subjectivity of qualitative judgments.

Stage 2: Batch CQ Template Instantiation Metrics Performance in Stage 2 was evaluated using fully automated, ground truth (GT)-referenced, and proxy-based metrics capturing correctness, completeness (total number of instantiated CQs by the LLM / total number of possible CQ instantiations), robustness (hallucination rate, prompt compliance,

and consistency), and efficiency of filled-in question and answer generation across multiple entities.

- **Slot-Level Exact Match (Slot Accuracy):** The proportion of individual variable slots (e.g., {Name}, answer fields) that exactly match the programmatic GT (a rule-based set of filled-in questions and answers generated programmatically, discussed further in LLM Experiment 2, Section 5.8) after normalisation.
- **Entity-Level Full Slot Match:** A binary metric indicating whether all required slots in a generated filled-in question-answer pair are correct for a given entity-template instance.
- **Semantic Similarity (SBERT Cosine Similarity):** Mean cosine similarity between GT and generated outputs, computed for both filled-in question-answer pairs and answers alone, following one-to-one alignment.
- **Semantic Similarity (BERTScore F1 [43]):** Token-level semantic overlap between generated and GT outputs, capturing lexical variation at the answer and question-answer level.
- **Strict Completeness:** The proportion of fillable template-entity pairs for which the LLM produced an output classified as exact or acceptable under calibrated semantic thresholds / total number of possible CQ instantiations.
- **Lenient Completeness:** As strict completeness, but including outputs rescued via lexical similarity rules and domain-specific exceptions.
- **Hallucination Rate:** The proportion of generated outputs flagged as unsupported, semantically incorrect, or non-compliant with output format constraints.
- **Prompt Compliance (Binary):** Whether generated outputs adhere to required structural constraints, including valid JSON formatting and absence of disallowed markup.
- **Consistency:** Variability of accuracy, completeness, and hallucination metrics across repeated runs with identical prompts, reported via standard deviation and confidence intervals.
- **Efficiency (Runtime):** Mean execution time per batch and per full run, used to assess practical feasibility within curator systems.

Ground Truth (GT)-Independent Validation Instrument To support operational settings in which a full programmatic Ground Truth (GT) is unavailable, a lightweight JSON Validation Proxy was defined as an auxiliary instrument. This classifier applies URI matching, slot presence checks, and lexical similarity rules to label outputs as exact, acceptable, or mismatched. Its performance was quantified against the GT-dependent semantic classifier using accuracy, precision, recall, and F1-score, using confusion matrices,

capturing its reliability as a first-pass validation mechanism during LLM Evaluation 2. It is further used as part of User Evaluation 3 and the Letter Sampo proof-of-concept to determine structural validity of LLM-generated filled-in questions and answers.

Semantic Evaluation Models and Calibration. Semantic similarity metrics employed pre-trained transformer models: `all-MiniLM-L6-v2` [119] and `all-mpnet-base-v2` [120] for SentenceBERT-based cosine similarity, and RoBERTa-Large [121] for BERTScore F1. Classification thresholds distinguishing exact, acceptable, and mismatched outputs were calibrated using a pooled, multi-LLM, manually labelled dataset and cross-validated macro-F1 optimisation to ensure task-specific validity.

Model Coverage. All instruments were applied across both closed- and open-source LLMs to enable analysis of trade-offs between accuracy, robustness, latency, and deployment constraints: GPT-5 [53], Gemini 2.0 Flash [54], DeepSeek-V3.1 [122], and Qwen-2.5-72B-Instruct [123].

Interpretive Scope. These instruments suggest system-level performance characteristics of LLMs within TMv1 rather than model benchmarking in isolation. Metrics are interpreted as indicators of practical suitability for curator-facing systems, not as claims of general-purpose model superiority.

5.2.3 Statistical Procedures Used In Relevant Evaluations

Statistical analysis procedures were selected based on data distribution characteristics, measurement scale, and experimental design.

Assumption Testing Normality of paired differences was assessed using the Shapiro-Wilk test [124] prior to selecting parametric or non-parametric statistical tests. This ensured that inferential procedures were aligned with the empirical distribution of the observed data.

Inferential Tests and Effect Sizes For normally distributed paired samples, paired t-tests were employed, with effect sizes reported using Cohen’s d [125]. Where normality assumptions were violated, Wilcoxon signed-rank tests [126] were used as a non-parametric alternative, with effect sizes reported using the rank-biserial correlation r . These complementary measures enabled both statistical significance and practical magnitude of effects to be assessed.

Multiplicity Correction To control for inflated Type I error rates arising from multiple hypothesis testing, Holm-Bonferroni correction [127] and Benjamini-Hochberg False Discovery Rate (FDR) [128] procedures were applied as appropriate, depending on whether the analysis prioritised family-wise error control or expected false discovery control.

Reliability Analysis Internal consistency of multi-item questionnaire instruments was assessed using Cronbach’s α . Where multiple raters were involved, inter-rater reliability was assessed using the Intraclass Correlation Coefficient (ICC) [118], providing an estimate of measurement stability and agreement.

5.3 Baseline System and Integration Context

User Evaluation 1 and User Evaluation 2 presented in this chapter were conducted using the VRTI Explorer [37], an existing web-based interface designed to support exploration of the VRTI KG. Although the VRTI Explorer evolved over the course of this research, from an early, minimal search-based interface to a more comprehensive exploration environment, its core role within this thesis remained consistent. Across all evaluations, the Explorer provided the underlying mechanism for accessing and navigating the VRTI KG, while Tús Maith was integrated as an additional feature layer intended to support initial exploration. Where changes to the Explorer materially affected the evaluation context, these are described explicitly in the corresponding evaluation sections.

All evaluations and experiments utilised the publicly deployed VRTI KG, rather than any subset or fork of it.

5.4 Domain Expert Context

During the course of the five evaluations presented below, domain experts, specifically historians from the VRTI project, were integral as both evaluators and participants. However, due to the specialised nature of the VRTI KG and the limited pool of suitably qualified historians, some degree of overlap between domain experts across evaluations was unavoidable. This overlap is made explicit in Table 5.2, which summarises the roles played by each domain expert in each evaluation.

Table 5.2: The role each of the three domain experts played in the evaluations they participated in.

Experiment	DE1 Role	DE2 Role	DE3 Role
LLM Experiment 1	CQ Template Evaluator		CQ Template Evaluator
User Evaluation 1	CuQA Generator		
User Evaluation 2	CuQA Generator	CuQA Generator	
User Evaluation 3		Participant	Participant

More specifically, the extent of the overlap per domain expert can be understood as the following:

DE1 was involved in LLM Experiment 1, User Evaluation 1, and User Evaluation 2. They did not participate in User Evaluation 3, the evaluation of TMv1 end-to-end. Their involvement therefore was limited to assessing abstract question artefacts (CQ Templates) or the curation of CuQAs rather than engaging with the system or interface under evaluation. In this context, DE1’s deep understanding of the VRTI KG and intended use-case can be considered advantageous, particularly for evaluating CQ templates, by enabling

informed judgments about coverage, relevance, and historical meaningfulness rather than introducing evaluative bias.

DE2 contributed to the generation of CuQAs used in User Evaluation 2 but had no prior involvement in CQ template evaluation or LLM-supported generation experiments. When participating in User Evaluation 3, DE2 encountered TMv1 for the first time and had no prior exposure to how it operated as a system. As such, while DE2 was familiar with the concept of CuQAs, their assessment of TMv1 reflects first-contact interaction with the system itself.

DE3 participated exclusively in LLM Experiment 1 as an evaluator of CQ templates and had no role in CuQA authoring or earlier user evaluations. Their involvement was limited to assessing abstract, decontextualised CQ templates rather than instantiated CuQAs or TMv1. Consequently, when DE3 later participated in User Evaluation 3, their prior familiarity was restricted to the conceptual design space of CQs rather than to TMv1 or its operationalisation.

Across all cases, no domain expert evaluated an artefact that they had previously designed, instantiated, or interacted with in its evaluated form. Prior involvement was confined to earlier, conceptually adjacent stages of the research pipeline, and did not include exposure to TMv1 as an end-to-end instantiation prior to its evaluation. This separation supports the validity of the reported findings.

It is important to note that the involvement of domain experts in authoring CuQAs without a curation system does not negate the motivation for the Tús Maith framework, or its instantiation, TMv1. While domain experts are capable of producing CuQAs, their contributions in User Evaluations 1 and 2 were intentionally limited in scope and intended to provide illustrative, rather than exhaustive, examples for evaluation purposes. These CuQAs were not designed to systematically explore the full semantic space of the VRTI KG.

Furthermore, as demonstrated in LLM Experiment 1, CQ templates authored by a single domain expert did not exhaustively capture the range of plausible CQ templates supported by the classes and properties of the VRTI KG. The use of LLM-supported generation increased coverage by surfacing additional, semantically valid CQ templates. As such, TMv1 is therefore not intended to replace domain expertise, but to aid it by supporting scalable generation and curation of CuQAs, enabling experts to focus on validation and refinement rather than on manual enumeration while maintaining strict oversight during the generation process.

5.5 User Evaluation 1: CuQAs as Entry Points

User Evaluation 1 was the first user evaluation undertaken. It was an exploratory study with the primary aim of determining whether CuQAs could function as exploratory entry points.

5.5.1 Exploratory Research Questions

At the time of this evaluation, the study was motivated by the following high-level hypothesis:

H1_UE1: The Tús Maith features* provide new users with a useful starting point for beginning exploration of the VRTI KG.

*Tús Maith subsequently becomes *just* CuQAs. However, at the time of the user evaluation, multiple features were considered.

Given the exploratory nature of this evaluation, this umbrella hypothesis was not tested directly. Instead, it was decomposed into a set of more specific exploratory research questions examining how perceived helpfulness manifests across different interface features and user backgrounds. The following exploratory research questions therefore are addressed by the study and the results informed subsequent design decisions.

RQ1_UE1: How do users perceive the helpfulness of Tús Maith features relative to other available interface features when beginning exploration of a KG?

RQ2_UE1: How does a user's familiarity with the contents of the KG relate to their perception of helpfulness of CuQAs as an entry point?

RQ3_UE1: Is providing different sets of CuQAs based on a user's background perceived as helpful?

RQ4_UE1: Are CuQAs perceived to give users a clearer idea of the information contained within the VRTI KG?

5.5.2 Experimental Setup

5.5.2.1 System Context

At the time of this evaluation, Tús Maith was conceived as a set of features comprising multiple entry-point mechanisms; namely, CuQAs, as well as commonly searched keywords, used in combination with a traditional search bar as seen in Figure 5.1 (the features in order identified by C, B, and A within the figure). Consequently, this evaluation did not isolate the effect of CuQAs in separation, but instead examined how CuQAs function relative to and in combination with other entry-point features within an early version of VRTI Explorer [37] (**RQ1_UE1**).

5.5.2.2 Features Under Examination

CuQAs CuQAs in this first iteration of design were a set of predefined questions created by a domain expert (a VRTI historian) to guide a new user in beginning their exploration of a KG. Users first selected a category (in this case, expertise level) and then were shown the corresponding set of CuQAs integrated as a drop-down menu. When the user clicked the "Generate Answer" button after having chosen their level of expertise and question, they triggered a predefined SPARQL query and were provided with both the answer to the question and additional information about the answer entity.

Figure 5.1: UE1 Implementation of Tús Maith.

Commonly Searched Keywords Commonly searched keywords are a set of search terms that have been frequently searched by users in the past. These keywords intend to provide a new user with another means of beginning their exploration by being able to see what others have been looking at within the KG.

5.5.2.3 Implementation of Features

For the purposes of the evaluation, five words were chosen by the author of this thesis to be the commonly searched keywords. These words were chosen pseudorandomly and aimed to highlight the different areas of the KG that were present at the time: People (2 terms), Occupation (1 term), Title (1 term), Area of Interest (1 term). They were displayed as clickable links that, when clicked, populated the search bar with the search term.

Regarding the CuQAs, a historian independent of Tús Maith (DE1) who is heavily involved in the creation of the VRTI KG was asked to curate fifteen questions that aimed to highlight the scope of the KG [129]. These questions were divided into three categories based on a user's knowledge of Irish history: beginner, five questions designed for users with little to no prior knowledge of Irish history, intermediate, five questions for users with some familiarity of Irish history, and expert, five questions for users with an advanced knowledge of Irish history. The categorisation was provided to answer **RQ3_UE1**, that users with different backgrounds may be interested in different aspects of the KG when beginning their exploration of it.

5.5.2.4 Participants and Procedure

Before participant recruitment, the evaluation underwent research ethics approval to ensure the evaluation was ethically sound and that the anonymous data that was gathered, adhered to Trinity College Dublin's strict safety and confidentiality standards. The full ethics application can be seen in Appendix A.4.1.

Sampling Strategy The study utilised convenience sampling as the aim was not to achieve demographic representativeness, but to gather initial insights from participants with varying degrees of familiarity with the domain and the KG.

Inclusion Criteria Participants were required to be over 18 years of age, capable of providing informed consent, and able to interact with the VRTI Explorer interface. No prior knowledge of the VRTI project or Irish history was required. Participants were recruited through academic networks, which took the form of an email providing potential participants with a brief overview of the evaluation, and a link to the questionnaire (see Appendix A.1.1). This resulted in twelve participants total with varying self-reported levels of familiarity with the VRTI project.

No explicit exclusion criteria were applied beyond failure to provide informed consent or to complete the questionnaire.

5.5.3 Evaluation Procedure

Before interacting with Tús Maith, the participants were asked to provide a score between one and five denoting their familiarity with the VRTI project (one being no familiarity and five being extreme familiarity). They were then provided with an image of the interface with the addition of Tús Maith (Figure 5.1) and the interface without (Figure 5.2) the addition of Tús Maith features and asked to choose which one they would find more appealing to begin their interaction with. Once they had selected their chosen version, they were required to interact directly with Tús Maith. Following this interaction, they were asked to rank the features in order of how helpful they found them when beginning their exploration of the KG. They were subsequently asked if they thought that being provided a different set of questions based on their level of experience with Irish history is helpful when beginning their exploration, whether the CuQAs helped them gain a clearer understanding of the types of information available within VRTI, if they found each feature intuitive, and if so, how intuitive on a scale of one to five.



Figure 5.2: VRTI Explorer without Tús Maith.

5.5.4 Data Collection Instruments

The evaluation consisted of an anonymous questionnaire that was filled out asynchronously. It gave participants a slightly more detailed explanation of the VRTI project and a high level overview of Tús Maith. Once participants consented to be a part of the evaluation, they were asked a series of both quantitative (ranking and rating using a Likert scale [130]) and qualitative (open-ended comments) questions. In order to complete the questionnaire,

participants were required to interact directly with Tús Maith as part of the VRTI exploration interface. The questionnaire can be seen in Table 5.3. This questionnaire consisted of custom questions rather than utilising existing usability questionnaire paradigms (e.g. PSSUQ [39, 40]). The two reasons for this were: (1) because the object of the evaluation was granular in nature: an attempt to understand how each feature aided users in comparison to the other as opposed to the usability of the overall system which is what such questionnaires have been designed for, and (2) the evaluation was exploratory in nature, it was not looking to exhaustively cover all aspects of usability (e.g. TAM [41]), but instead identify next directions.

Table 5.3: Questionnaire Questions for User Evaluation 1

ID	Question
1	How familiar are you with the Virtual Record Treasury of Ireland project?
2	Which version would you find more appealing to start your interaction with, Version A (the VRTI webpage without Tús Maith) or Version B (the VRTI webpage with Tús Maith)?
3	How much do you like your chosen option?
4	Rank the three features marked A, B, and C in order of how helpful you found them in beginning your exploration webpage. [Most Helpful]
5	Rank the three features marked A, B, and C in order of how helpful you found them in beginning your exploration webpage. [Somewhat Helpful]
6	Rank the three features marked A, B, and C in order of how helpful you found them in beginning your exploration webpage. [Least Helpful]
7	Did you find that the curated questions aligned with the level of experience with Irish history you indicated on the webpage?
8	Do you think that being provided a different set of questions based upon your level of experience of Irish history is helpful when beginning your exploration?
9	Did the curated questions help you gain a clearer understanding of the types of information available within VRTI?
10	Did you find the curated questions intuitive to use?
11	If yes, please rate between 1 and 5 denoting how intuitive you found them.
12	Do you have any comments or suggestions? If so, please write them here.
13	Did you find the commonly searched keywords intuitive to use?
14	If yes, please rate between 1 and 5 denoting how intuitive you found them.
15	Did you find the search bar intuitive to use?
16	If yes, please rate between 1 and 5 denoting how intuitive you found them.

5.5.5 Results and Design Implications

This section outlines the results of the study and the implications of the results for the next iteration of the development of Tús Maith. A core component of this evaluation was the consideration of the context in which the users were interacting with the features,

specifically, their level of background knowledge of the VRTI project, as the IEP was hypothesised to most affect those with limited to no familiarity of a KG’s contents. This familiarity spread is presented also within the results of the evaluation below.

5.5.5.1 Presentation of Tús Maith

When presented with the screenshots shown in Figure 5.1 and Figure 5.2, before they interacted with Tús Maith, and they were asked which they preferred, nine participants out of the twelve chose the version without Tús Maith. One possible explanation is that the version with Tús Maith appeared more cluttered; however, this remains speculative as participants were not asked to justify their choice.

5.5.5.2 Ranking of Features

Once the participants had chosen their preferred version, they were then given the opportunity to interact with the Tús Maith implementation and were encouraged to use all three features: CuQAs, commonly searched keywords, and search. After familiarising themselves with Tús Maith, they were then asked to rank the three features in order of what they found most to least helpful when beginning their exploration of the VRTI KG, the results of which can be seen in Figure 5.3. It was interesting to note that CuQAs and search were both ranked most often as the most helpful and least helpful features while commonly searched keywords was generally ranked as somewhat helpful. This indicated that while commonly searched keywords may not be of the most use when beginning exploration of an unfamiliar KG, it was valuable as a supporting feature. As it did not appear to address the onset of exploration which is the focus of this thesis, this motivated the scaling back of what Tús Maith consisted of to just CuQAs in all further evaluations.

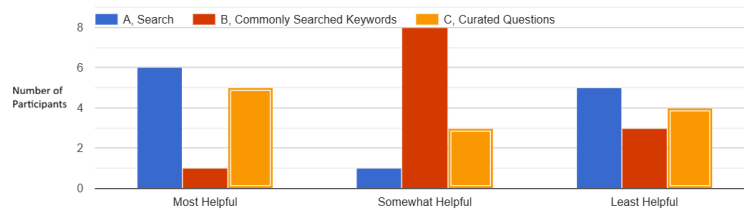


Figure 5.3: Feature Helpfulness Ranking.

Scores of familiarity of participants with VRTI were distributed as follows (1 being not familiar at all, 5 being extremely familiar): score 1 - 2 responses, score 2 - 1 response, score 3 - 1 response, score 4 - 2 responses, and score 5 - 6 responses. An interesting trend emerged when analysing how participants’ familiarity with the VRTI project influenced their ranking of the features. As illustrated in Figure 5.4, it is possible to see that those with low familiarity of the project (a familiarity score of three or less), predominantly chose CuQAs as the feature they found the most helpful. This is in contrast to the participants with high familiarity with the project (a familiarity score of four or higher), who overwhelmingly preferred search as the feature that they would choose to begin their exploration of the VRTI KG. These results suggest a distinction in perceived usefulness: participants with

little to no prior knowledge of the KG more frequently reported CuQAs as helpful starting points. However, for users already acquainted with the KG, CuQAs were less beneficial, as they were more likely to rely on the search bar to begin their navigation of the KG. This observation was helpful with respect to the primary objective of Tús Maith, which is to assist new users in beginning their exploration of an unfamiliar KG. It is worth noting however, that despite the strong preference for the search bar among participants with high familiarity, two participants still ranked CuQAs as the most helpful feature. This suggested that while individuals with prior knowledge of the KG are more likely to prefer the search bar, there still was a subset of such users who found the CuQAs valuable. This nuance highlighted the potential for CuQAs to serve a broader range of users, even users with some familiarity may still benefit from guided exploration (although this was not the primary focus of the evaluation).

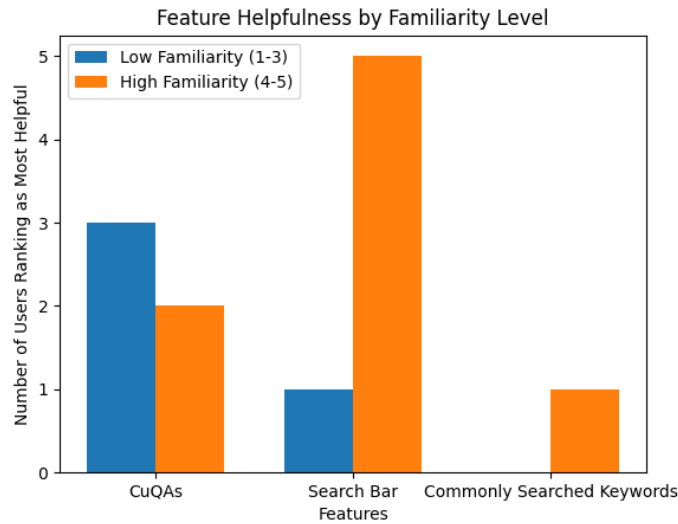


Figure 5.4: Feature Helpfulness by VRTI Familiarity Level - Most Helpful

This observation was reinforced by Figure 5.5, showing that none of the low-familiarity participants ranked CuQAs as least helpful, compared to four with high familiarity.

5.5.5.3 Feature Intuitiveness

In general, the participants found all three features intuitive to use, with ten of the twelve participants finding the commonly searched keywords and search intuitive and eleven of the twelve finding the CuQAs intuitive. This was a positive outcome indicating that the features as currently designed did not pose a significant usability challenge to the participants. For features rated as intuitive, participants rated each feature’s intuitiveness on a 1-5 Likert scale [130] with one being only a little intuitive and five being extremely intuitive. Overall, search was rated 4.2/5, CuQAs 4/5, and commonly searched keywords 3.5/5. Again, generally a positive result for search and the CuQAs. It was noted in the comment section that the commonly searched keywords did not feel as intuitive as the other features due to the fact that while each keyword looked like a link to a webpage, when the keyword was clicked, the search bar was populated instead. This functionality

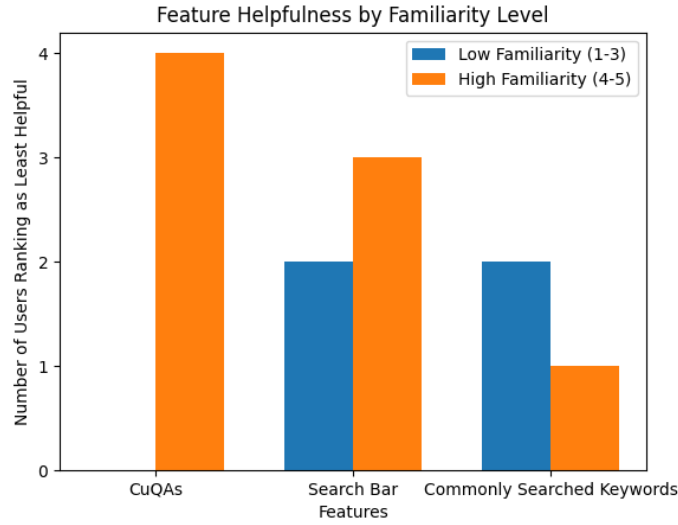


Figure 5.5: Feature Helpfulness by Familiarity Level - Least Helpful

was intentionally designed to emphasise the connection between the keywords and the search bar. However, user feedback revealed that this approach caused confusion. Another comment highlighted confusion over the CuQAs in the curated section, which gave the impression that the search bar would support similar queries, even though it only accepted search terms. To address this, the next iteration of Tús Maith as seen in User Evaluation 2, Section 5.6, separated the CuQAs from the search bar to avoid confusion and to be less overwhelming for new users.

5.5.5.4 Analysis of CuQAs

The participants were also asked whether or not they found that the CuQAs with which they were provided aligned with the level of expertise with Irish history that they had indicated. Most agreed that the questions aligned well with their expertise levels (ten out of twelve) which indicated that the domain expert crafting of questions tailored to an individual's level of expertise was generally perceived to be accurate. This result indicated the potential importance of involving a domain expert in the design of CuQAs. Participants were additionally questioned as to whether providing a different set of questions according to an individual's level of experience with Irish history was helpful when beginning their exploration. While the majority (eight out of twelve) agreed that it was helpful, in the next iteration of Tús Maith, the CuQAs would instead be categorised by category of information within the KG. This is because a qualitative trend emerged from participant comments that indicated that being able to explore themes and attributes would potentially be of more value to someone when beginning their exploration of the KG.

One additional notable outcome of the evaluation with respect to the CuQAs was that eleven out of twelve participants agreed the questions provided them with a clearer understanding of the types of information available within the KG. Participants described the CuQAs as "great overall", that they were "extremely valuable" in allowing the participant to "see what kind of data was present", that they "served as a good starting point", and

provided "a straightforward way to discover different types of data available in the VRTI KG". This feedback motivated further inquiry into exactly how effective CuQAs were in addressing the characteristics of the IEP leading to User Evaluation 2 (Section 5.6).

5.5.6 Summary of Key Findings

Firstly, in looking at the exploratory research question, **RQ1_UE1**: "How do users perceive the helpfulness of CuQAs relative to other available interface features when beginning exploration of a KG?" the results indicated that the helpfulness of CuQAs is dependent on the level of familiarity of the user. They appeared to be particularly effective for users with a low familiarity with the KG when beginning their exploration while in contrast, users with high familiarity preferred the search bar, as it allowed them to directly access the information they sought. This insight ties directly into **RQ2_UE1**: "How does a user's familiarity with the contents of the KG relate to their perceived helpfulness of CuQAs as an entry point?" It appears that a user's familiarity and their perception of CuQAs as entry points are closely linked. This exploratory evidence showed early on that CuQAs were worth pursuing as a means of addressing the IEP which affects lay users at the onset of exploring an unfamiliar KG.

While results showed that the CuQAs provided tended to align with the user's self-reported background, the qualitative evidence did not sufficiently support continuing to use level of familiarity as a means of creating set of CuQAs addressing **RQ3_UE1**: "Is providing different sets of CuQAs based on a user's background perceived as helpful?" Instead, based on the free-text comments, the decision was made to split the sets of CuQAs by primary category of information contained within the KG for all further evaluations. The categories being: Person and Place.

Finally, the evidence gathered indicated that the majority of participants perceived CuQAs as helping them gain a clearer understanding of the types of information contained within the VRTI KG directly addressing **RQ4_UE1**: Are CuQAs perceived to give users a clearer idea of the information contained within the VRTI KG? Once again, motivating further investigation into them via a more targeted user study as detailed in Section 5.6.

5.6 User Evaluation 2: Addressing the Initial Exploration Problem

After User Evaluation 1, Section 5.5, indicated the promise of CuQAs as entry points for lay people unfamiliar with the VRTI KG, the next question to be answered was: to what extent do CuQAs address the barriers that characterise the Initial Exploration Problem (IEP)? User Evaluation 2 therefore aimed to understand this through an asynchronous online questionnaire with pre- and post-questions based on the participant's use of the VRTI Explorer without and with CuQAs.

As defined in Chapters 1 and 3, the framing of the IEP consists of three interrelated barriers - scope uncertainty, ontology opacity, and query incapacity. However, as a result of the design of the VRTI Explorer and CuQAs, not all three barriers were directly observable

at interaction level in this evaluation. Specifically, query incapacity could not be meaningfully measured directly, as participants were neither exposed to nor required to engage with formal query languages (e.g. SPARQL). Query formulation was fully abstracted behind keyword-based search, other interface features such as a people/place randomiser, and the CuQAs themselves. As a result, participants were unable to experience or articulate difficulties related to query construction, syntax, or formal expressivity.

Rather than attempting to measure a latent barrier that was intentionally hidden from users (query incapacity), this evaluation instead measured orientation, an emergent user state reflecting a participant’s perceived ability to form a meaningful starting point for exploration as surfaced in Chapter 3. The difficulty of achieving such orientation in KG exploration has been noted in prior work [131] and is treated here as reflecting the extent to which the convergent barriers of the IEP are experienced as sufficiently mitigated and viable entry points are perceived to exist. It is not binary and does not require the elimination of uncertainty. Rather, it emerges once IEP barriers are mitigated below a perceptual threshold sufficient to support confident forward movement. Baseline orientation may arise from domain familiarity and interface familiarity, but epistemic scaffolds such as CuQA entry points deepen this orientation by making the graph’s underlying structure and potential explicit. In other words, baseline orientation can occur as a reflection of interface-level navigability, whereas CuQAs aim to support structural orientation by making the KG’s underlying entity types, relations, and epistemic scope more legible.

More formally, the hypothesis and research question are presented below.

5.6.1 Hypothesis and Research Question

H1_UE2: CuQAs are perceived to address the observable barriers that characterise the IEP for lay users unfamiliar with the VRTI KG.

H1_UE2 can be better understood by the following research question:

RQ1_UE2: To what extent do CuQAs address perceived scope uncertainty, perceived ontology opacity, and improve perceived orientation for lay users unfamiliar with the VRTI KG?

5.6.2 Experimental Setup

5.6.2.1 System Context

During the course of the research conducted as part of this thesis, the VRTI Explorer [37] underwent significant development. As a result, the CuQAs were integrated into an interface that now provided the following features: not only a search bar with advanced filtering capabilities, but people and place perspectives, a randomiser surfacing random people or places contained within the VRTI KG, and the provision of groups of the people found within a set of historian curated collections. It also was visually different to the version used as part of User Evaluation 1, Section 5.5. However, the overall aim of the VRTI Explorer remained consistent: to provide lay users with a way to interact with the VRTI KG.

5.6.2.2 Implementation of CuQAs

As part of the evaluation, the participants were required to interact with the public VRTI Explorer *without* CuQAs¹ and a prototype version of the VRTI Explorer *with* CuQAs². The way in which CuQAs were integrated as part of the VRTI Explorer can be seen in Figure 5.6. For the purpose of the evaluation, the user-facing name for CuQAs was "Featured Questions". Conceptually and functionally however these questions and answers were still CuQAs.

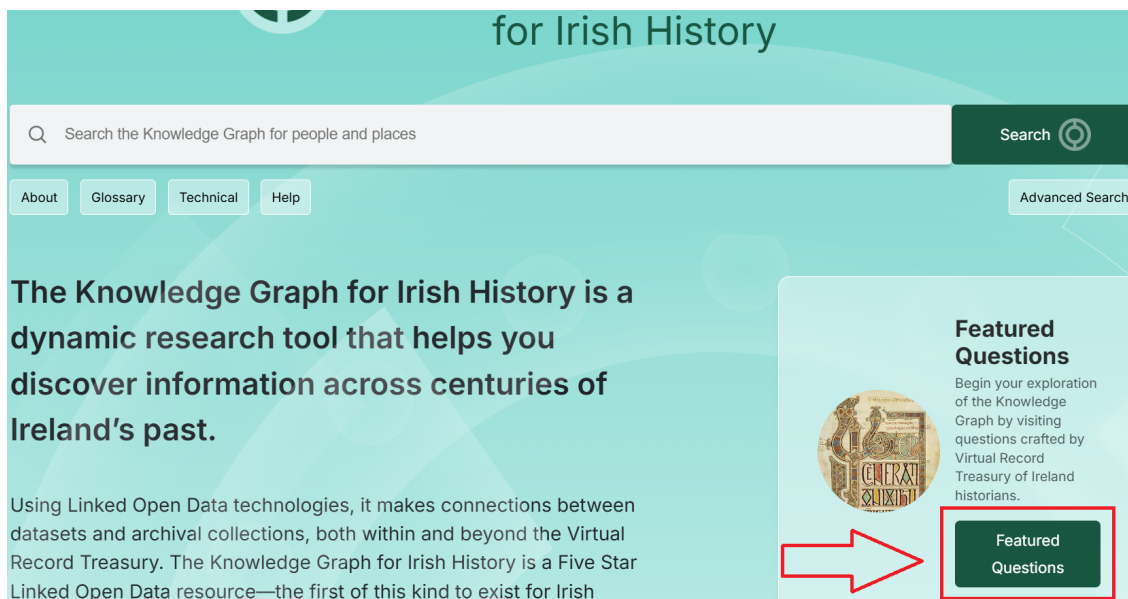


Figure 5.6: CuQAs implemented within the VRTI Explorer as Featured Questions.

As the development of the TMv1 prototype had not yet been completed end-to-end, domain expert, DE1, and a second domain expert also heavily involved in the VRTI project, DE2, were tasked with coming up with 3 - 5 CuQAs related to people and 3 - 5 CuQAs related to places that could be used as part of the evaluation implementation. Splitting the question sets into the two primary categories of information found in the VRTI KG rather than categorising by a user's background follows from the qualitative evidence obtained via User Evaluation 1, Section 5.5 and is analogous to a temporally bounded version of the idea of "conceptlenses" proposed by [132]. The domain experts were instructed that the CuQAs needed to be related to (but not exclusively) content found within the KG as it currently exists and, for the answers to include entities that are as rich and interesting as possible. The email sent to the domain experts can be found in Appendix A.1.3. The CuQAs that DE1 and DE2 created can be seen in Table 5.4.

¹<https://kg.virtualtreasury.ie/>

²<https://mcnamacl.pythonanywhere.com/>

Table 5.4: Example featured questions categorised by type and domain.

Type	Category	Example Question & Answer
Simple	Place	Q: What can I discover about life on the Aran Islands before the Famine (1845-52)?
		A: https://kg.virtualtreasury.ie/entity-card/place/present-day/townland/Killeany/v13dqr4
		You can discover life on the Aran Islands before the Famine (1845-52) by clicking on the search result below!
Interface Dependent	Place	Q: I'm interested in local history. What kind of information might I find?
		A: Jump to featured places carousel: https://kg.virtualtreasury.ie/place-homepage
		You can find local places of interest here. Click on a place to learn more about it!
Interface Dependent	Place	Q: Almost all Ireland's pre-Famine Census records were destroyed. Are there surviving records for my county?
		A: https://kg.virtualtreasury.ie/collection-map/census-gleanings
		You can search for your county in the search box below. Zoom in to see specific results for townlands.
Simple	People	Q: Who might have been involved in the 1798 Rebellion?
		A: https://kg.virtualtreasury.ie/search-results?People:collection=Age+of+Revolution&People:sortBy=Surname&People:sortOrder=ASC
		You can see the people who might have been involved in the 1798 Rebellion in the search results below. Click any search result below to find out more about that person!
Interface Dependent	People	Q: How can I explore Irish lives outside Ireland?
		A: Jump to map of birth and death location on 'people' homepage: https://kg.virtualtreasury.ie/people-homepage
		You can explore the map below by using your mouse to drag and zoom anywhere in the world!
Complex	People	Q: Strongbow led the Norman invasion of Ireland in 1169. Which woman inherited his vast lands and wealth?
		A: https://kg.virtualtreasury.ie/entity-card/person/Marshal_Isabella_c13/v154qmw
		Isabella Marshal inherited Strongbow's vast lands and wealth. You can find out more about her by clicking on the search result below!

5.6.2.3 Participants and Procedure

The evaluation design underwent full research ethics approval to ensure that it adhered to Trinity College Dublin's strict ethical standards for research as can be seen in Appendix A.4.2.

Sampling Strategy The population for consideration was university-level students with an interest in history taking a module taught by Professors working with or adjacent to the VRTI project. From that population, voluntary response sampling was used to recruit participants who would be willing to participate in the study.

Inclusion Criteria The inclusion criteria were as follows: students with an interest in Irish history taking a relevant module run by a Professor working with or adjacent to the VRTI Project. The Professors were provided with the link to the questionnaire (described in Section 5.6.2.4) and asked to either give their students time at the end of a lesson to complete the evaluation or ask the students to do the evaluation at home in a specific time frame. In this manner, two different VRTI related history Professors presented the evaluation to their students, resulting in 35 participants from three student cohorts ranging from undergraduate to MPhil.

Evaluation Procedure Pre-interaction with the VRTI Explorer without and with CuQAs, participants were given a brief introduction to Tús Maith, namely, that it was designed to help new users begin exploration of the VRTI KG and that this was operationalised via a set of featured questions (CuQAs). Then participants were asked two background questions to obtain contextual information for their results - whether or not they had used the VRTI Explorer before today, and how familiar they were with the VRTI project.

Upon answering those questions, they were instructed to use the VRTI Explorer without featured questions (CuQAs) by interacting with the public facing VRTI Explorer interface for a few minutes, and then to return to the questionnaire. On their return, they were asked the pre-use of CuQAs orientation question of how confident they felt in knowing where to start exploration of the VRTI KG. Following this, the participants were asked to carry out 8 loosely-defined tasks which can be viewed in Table 5.5 that required utilisation of the featured questions (CuQAs). Once they had completed this, they were asked the set of post-CuQA interaction questions.

Table 5.5: Featured Questions (CuQA) Interaction Tasks.

Step	Action
1	Begin by clicking the "Featured Questions" button from the right side panel.
2	Then click "person" to view the person related featured questions. Choose one that you find interesting or would naturally want to know more about.
3	Explore the results page with the answer that loads. If you find an information thread that interests you, feel free to continue exploring by clicking on the available filters, links, etc. as you wish.
4	When you're ready to move on, return to the homepage by clicking the Knowledge Graph logo in the top left corner.

Table 5.5: Featured Questions (CuQA) Interaction Tasks. (continued)

Step	Action
5	Click the "Featured Questions" button once again.
6	This time select "place" to view place-related questions.
7	In the same way as you did for person, choose one that catches your interest.
8	Again, when you get brought to the results page with the answer, explore as much or as little as you like.

5.6.2.4 Data Collection Instruments

Similar to User Evaluation 1, Section 5.5, participants were provided with a questionnaire (see Table 5.6) via Google forms that was to be filled out asynchronously. The questionnaire presented CuQAs as featured questions to the participants and explained that it was a new feature designed to provide exploratory entry points to new users of the VRTI Explorer. While the questionnaire was purpose-built to evaluate barriers associated with the IEP, some items were conceptually inspired by established usability and technology acceptance instruments, including PSSUQ [39], SUS [40], and TAM [41]. These were complemented by custom questions tailored to exploratory KG interaction, along with an open-ended qualitative question. The questionnaire questions and their category can be seen in Table 5.6. Aside from the yes/no question and qualitative open-ended comment question, each question required a response on a 1 - 5 Likert scale [130], where 1 was "not at all" and 5 was "very".

Table 5.6: Questionnaire items and their corresponding categories

Question/Statement	Category
Have you used the VRTI Explorer before today?	Contextual
How familiar are you with the VRTI project?	Contextual
I felt confident in knowing where to start exploring the VRTI dataset.	Orientation (Pre)
Which featured question for person did you select?	Behavioural
Which featured question for place did you select?	Behavioural
I found it easy to get started exploring the VRTI dataset using the featured questions.	Orientation (Post)
Without the featured questions, I found it more difficult to know where to begin exploring the VRTI dataset.	Counterfactual
The featured questions made me feel that I had a clearer idea of the types of information I can find in the VRTI dataset.	Scope Uncertainty
The featured questions gave me an idea of the range of categories of information available in the VRTI dataset.	Scope Uncertainty

Continued on next page

Table 5.6 – continued from previous page

Question/Statement	Category
The featured questions made the VRTI dataset feel more approachable.	Ontology Opacity
The featured questions led me to information I wouldn't have considered otherwise.	Scope Uncertainty
Following a featured question helped lead me to an area of the VRTI dataset that I was curious about.	Exploration Curiosity
The featured questions made the dataset feel easier to explore.	Ontology Opacity
The featured questions helped me begin exploring without needing detailed background knowledge of the dataset.	Ontology Opacity
Using the featured questions helped me identify meaningful next directions to continue exploring the dataset.	Orientation (Post)
After using the featured questions, I felt ready to continue exploring on my own.	Orientation (Post)
I would use featured questions again to help me explore the VRTI dataset.	Reuse
Please provide any comments you have about featured questions and/or your experience using them in the text box below. (Do not input any identifiable information about yourself)	Qualitative

More specifically, there was one question relating to orientation pre-interaction with the VRTI Explorer with CuQAs, and three questions relating to orientation post-interaction. There were also three questions assigned per the two observable IEP barriers.

5.6.3 Statistical Approach Used

The evaluation adopted a mixed parametric and non-parametric statistical approach to analyse the Likert-scale questionnaire data. The internal reliability of the constructs being measured was assessed, as was the magnitude and robustness of the observed effects.

5.6.3.1 Construct Formation and Assessment of Reliability

Three post-interaction constructs were used to analyse the perceived effect of CuQAs on the observable barriers of the IEP (Scope Uncertainty, Ontology Opacity) in the context of the VRTI Explorer, and the emergent effect of addressing those barriers (Orientation). Composite scores were calculated as the mean of the ratings provided for each of the three questions associated with the construct.

Internal reliability was assessed using Cronbach's α [133] which evaluates the degree to which items within each construct measure the underlying concept. Thresholds for Cronbach's α were interpreted following guidance on internal consistency reliability, whereby values between 0.60 and 0.70 are considered acceptable in exploratory research [134], and

values above 0.70 indicate at least moderate to good reliability. This interpretation is consistent with work demonstrating that Cronbach’s α provides a conservative lower-bound estimate of true construct reliability and may underestimate reliability when its assumptions are violated, particularly in short multi-item scales [135, 136].

5.6.3.2 Sample Size Justification and Statistical Power

An a priori power analysis was conducted using G*Power 3.1 [137] for a within-subject paired design with a one-tailed test. Using the Wilcoxon signed-rank test (matched pairs) with a normal approximation, assuming a medium effect size ($d_z = 0.6$), $\alpha = 0.05$, and statistical power ($1 - \beta = 0.95$), the required sample size was $n \geq 33$. As the final sample size exceeded this threshold ($n = 35$), the study was considered adequately powered to detect effects of practical significance.

5.6.3.3 Distributional Assumptions and Robustness Considerations

All questionnaire items for the constructs were measured on a 5-point Likert scale and are therefore formally ordinal. However, given the sample size ($n = 35$), and given the composite construction, the Central Limit Theorem (CLT) [138] supports robustness of parametric tests applied to mean-based composite scores. While that is the case, ensuring robustness, normality of the composite distributions and difference scores was also assessed using the Shapiro-Wilk [124] test. Where deviations from normality were observed, parametric tests were retained for comparability, but paired or one-sample non-parametric alternatives were also reported to ensure statistical conclusions did not depend on distributional assumptions.

5.6.3.4 Emergent Construct Analysis: Pre-Post Orientation Change

Orientation was measured as a within-subject comparison between the pre-interaction orientation item and the post-interaction orientation composite score. A paired-samples t-test was used to test for a mean difference in perceived orientation before and after use of featured questions (CuQAs). In line with also reporting non-parametric test results, a Wilcoxon signed-rank test was also conducted to evaluate the median differences.

For within-subject comparisons, effect sizes are reported as Cohen’s d_z , defined as the mean of the within-participant difference scores divided by their standard deviation ($d_z = \bar{D}/s_D$), which is equivalent to $d_z = t/\sqrt{n}$ for paired-samples t-tests.

For the Wilcoxon test, effect size r was computed using the standard normal approximation ($r = Z/\sqrt{N}$), where N is the number of non-zero paired differences. Directionality was determined from the sign of the standardised test statistic.

5.6.3.5 Observable IEP Barriers Analysis: Scope Uncertainty and Ontology Opacity vs. Neutral

To evaluate whether CuQAs were perceived to meaningfully address scope uncertainty and ontology opacity, post-interaction composite scores for both were tested against a neutral

reference point of 3, corresponding to the midpoint ("neither agree nor disagree") of the 5-point Likert scale.

One-sample t-tests were used to assess whether mean composite scores differed significantly from this neutral baseline. As above, Wilcoxon signed-rank tests were also used to test whether median scores differed from the same reference point.

Effect sizes for one-sample parametric tests are reported using Cohen's d , computed as the mean deviation from the neutral reference value divided by the sample standard deviation. For non-parametric tests, effect size r is reported for the Wilcoxon signed-rank test.

5.6.3.6 Control for Multiple Comparisons

In order to control for the risk of a Type I error [139] in evaluating whether CuQAs addressed a distinct IEP barrier or emergent outcome, Holm-Bonferroni correction [127] was applied across the three constructs. For each construct, parametric and non-parametric tests were treated as complementary robustness checks of the same underlying hypothesis rather than as independent inferential claims, and so multiplicity correction was applied at the level of constructs rather than individual test statistics. More specifically, Holm-Bonferroni correction was applied to the t-test and Wilcoxon p-values.

5.6.3.7 Qualitative Data Integration

Open-ended responses were reported to contextualise and interpret the quantitative results, particularly in cases where statistical outcomes indicated weak or non-significant effects. This triangulation supports construct validity by identifying interface-level factors and user perceptions that may explain observed patterns in the quantitative data.

5.6.3.8 Summary of Analytical Rationale

This combined analytical strategy was designed to ensure:

- Construct validity through internal reliability assessment and item diagnostics,
- Statistical robustness through complementary parametric and non-parametric testing,
- Practical relevance through systematic reporting of effect sizes,
- Error control through family-wise multiplicity correction, and
- Interpretive depth through qualitative triangulation.

5.6.4 Results and Implication for CuQAs as Mitigation of the IEP

This section presents the results of the study and outlines how the results can be understood with respect to CuQAs being perceived to mitigate the IEP.

5.6.4.1 Background Information on Participants

Figure 5.7 presents the breakdown of participants who had and had not used the VRTI Explorer prior to the evaluation. As is evident from the chart, most participants had not used the VRTI Explorer prior to the evaluation (88.6% or 31/35).

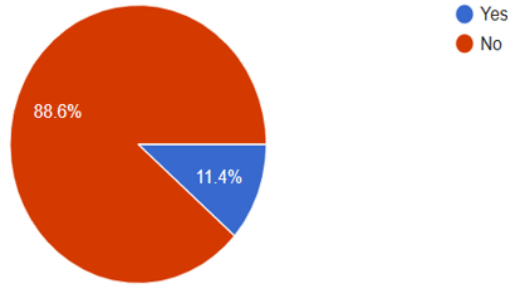


Figure 5.7: Piechart presenting the number of participants who had/had not used the VRTI Explorer prior to the evaluation (n=35).

Most participants had limited familiarity (≤ 2) of the VRTI project (71.4% or 25/35) as is visible in Figure 5.8.

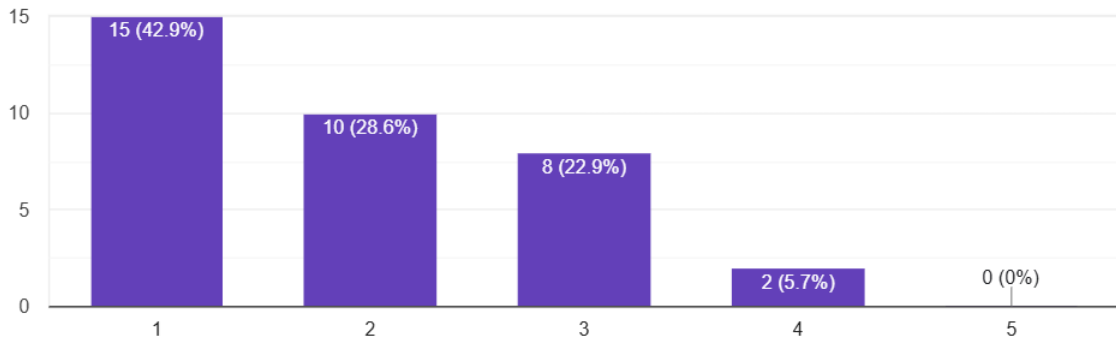


Figure 5.8: Participant's self-reported familiarity with the VRTI project (n=35) where 1="Not familiar at all" and 5="Extremely familiar".

5.6.4.2 Pre-Use of Featured Questions (CuQAs)

As seen in Figure 5.9, participants reported relatively high baseline levels of perceived confidence in where to start when exploring the VRTI dataset (mean = 3.69, median = 4.0). This elevated baseline introduced a potential ceiling effect that limited measurable improvement.

While the questionnaire did not include items explicitly questioning the reasons for initial confidence, it is possible to speculate regarding reasons why. Firstly, the public VRTI Explorer without featured questions (CuQAs) is well designed, providing a range of well-designed exploratory affordances including person and place perspectives, and a randomiser that surfaces entities within the KG. Secondly, the participant pool consisted of students with an academic interest in history, and thus likely possessed sufficient domain knowledge

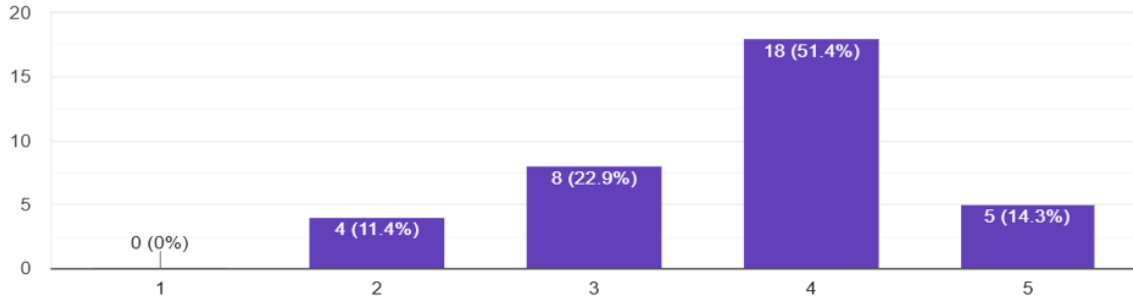


Figure 5.9: Participants' response to "I felt confident knowing where to start exploring the VRTI dataset" (n=35) where 1="Not confident at all" and 5="Extremely confident".

to formulate initial exploration strategies. This potentially reduced the perceived initial orientation problem.

Once again, the above is speculative as there was no question in the questionnaire that specifically probed the reasoning behind a participant choosing their rating. However, the qualitative comments provide convergent support for these interpretations. Several participants explicitly described the interface as "easy to navigate" and "well designed", and highlighted the randomiser as "incredibly helpful and entertaining" and effective for "anyone [to] find information regardless of their background knowledge". Others noted that even when moments of confusion occurred, they were readily resolved through existing interface features, such as filters and navigation controls. Collectively, these remarks suggest that the baseline KG explorer interface already afforded some viable entry points into the KG, which may have mitigated perceived orientation-related difficulties prior to the introduction of featured questions and contributed to the observed high pre-use confidence levels.

Something to note is that participants' qualitative feedback primarily referenced ease of navigation and interface clarity rather than understanding of what the KG could semantically answer. Further, while the randomiser surfaces entities and reduces initial interaction friction, it does not expose the semantic scope of the KG. As such, it supports entry into the interface without necessarily resolving scope uncertainty or ontology opacity.

5.6.4.3 Orientation Analysis: Pre vs. Post Use of Featured Questions (CuQAs)

The post-orientation construct demonstrated acceptable internal reliability (Cronbach's $\alpha = 0.63$) in exploratory contexts. Shapiro-Wilk analysis indicated that the data did not significantly deviate from normality ($p = 0.346$). However, as outlined in Section 5.6.3, both parametric and non-parametric tests were conducted over the results.

As seen in Table 5.7 neither test indicated a statistically significant change in perceived orientation following the use of featured questions (paired t-test: $p = 0.206$; Wilcoxon signed-rank: $p = 0.363$). The mean increase in orientation was small (mean difference 0.21 points, Cohen's $d_z = 0.218$), while the median change was 0.00 indicating at least half of participants reported no improvement or a decrease in perceived orientation.

This pattern is consistent with the high pre-use baseline observed in the pre-use of featured questions (CuQAs) orientation item seen above. The limited scope for upward movement suggests that the absence of a statistically and practically meaningful effect is likely due to a ceiling effect as opposed to a lack of utility of featured questions (CuQAs). In other words, participants may have already felt sufficiently orientated by the existing interface and their own domain familiarity, leaving little room for measurable gains. Furthermore, the study was powered to detect medium within-subject effects ($d_z = 0.6$); smaller effects, such as those observed for orientation, may therefore not have been detectable.

Table 5.7: Orientation: Pre- vs. Post-Use of Featured Questions (CuQAs)

Metric	Value
Sample size (n)	35
Pre-use Mean (SD)	3.69 (0.87)
Post-use Mean (SD)	3.90 (0.73)
Mean Difference (Post–Pre)	0.21
Median Difference	0.00
Cronbach’s α (Post Composite)	0.63 (acceptable)
Item-total correlations (r range)	0.53 – 0.87
Paired t -test	$t(34) = 1.29$
p -value (raw)	0.206
Cohen’s d_z	0.218 (small)
95% CI (Mean Diff.)	[–0.12, 0.54]
Holm-Bonferroni $p_{\text{corrected}}$	0.206
Significance ($\alpha = 0.05$)	Not significant
Wilcoxon Signed-Rank	$W = 175.5$
p -value (raw)	0.363
Effect size r	-0.169 (small)
95% Bootstrap CI (Median Diff.)	[–0.33, 0.67]
Holm-Bonferroni $p_{\text{corrected}}$	0.363
Significance ($\alpha = 0.05$)	Not significant

Orientation is conceptualised as an emergent, user-perceived state resulting from the perceived mitigation of scope uncertainty, ontology opacity, and interaction friction. However, such perceived mitigation may occur via domain familiarity or interface abstraction rather than structural transparency of the underlying KG. As a result, users may report confidence in knowing where to begin while not yet having awareness of the semantic scope of the KG or the range of questions it can meaningfully answer, meaning perceived orientation may reflect navigational entry confidence rather than a more comprehensive epistemic orientation. CuQAs are introduced as a direct structural intervention capable of addressing all three barriers, independent of user expertise and interface richness. Thus providing semantically meaningful entry points that expose KG answerability rather than simply supporting navigation. A deeper orientation to a KG does not consist solely in knowing how to navigate its interface, but in understanding what kinds of questions its structure enables (Belkin’s Anomalous State of Knowledge, ASK – “users don’t know what they don’t know”). Without such understanding through, for example, the provision of abstrac-

tion mechanisms such as CuQAs, the graph risks being experienced as little more than a searchable collection of records rather than a rich, interconnected semantic web.

5.6.4.4 Scope Uncertainty Analysis: Post Use of Featured Questions (CuQAs)

The scope uncertainty construct exhibited good internal reliability (Cronbach's $\alpha = 0.74$), indicating that the composite coherently captured participants' perceptions of the breadth and discoverability of the KG content. Shapiro-Wilk analysis indicated that the data significantly deviated from normality ($p = 0.0035$). However, as with orientation, both parametric and non-parametric tests were conducted for completeness.

Unlike orientation, scope uncertainty was not measured prior to interaction and therefore could not be analysed as a within-subject change. Instead, the post interaction composite score was evaluated against the neutral midpoint of the Likert scale (3) corresponding to "neither agree nor disagree".

The one-sample t-test indicated that the mean composite score (mean = 4.01, 95% CI [3.74, 4.28]) was significantly greater than the neutral reference point ($p \leq 0.001$) with a very large effect size (Cohen's $d = 1.26$). The non-parametric Wilcoxon signed-rank test corroborated this result, yielding a median of 4.0 and a large effect size ($r = -0.781$, $p \leq 0.001$). The results can be seen in the Table 5.8.

Following Holm-Bonferroni correction for multiple comparisons across the IEP barriers and emergent outcome, both results remained statistically significant ($p_{\text{corrected}} \leq 0.001$). These findings provide strong evidence that featured questions (CuQAs) were associated with positive perceptions of understanding of the scope of the KG when compared to neutral by helping participants form a clearer understanding of the breadth and types of information available within the KG.

Table 5.8: Scope Uncertainty: Post-Use of Featured Questions (CuQAs) vs. Neutral

Metric	Value
Sample size (n)	35
Mean (SD)	4.01 (0.80)
Median	4.00
Neutral Reference	3.00
Cronbach's α	0.74 (good)
Item-total correlations (r range)	0.79 – 0.85
One-sample t -test	$t(34) = 7.48$
p -value (raw)	< 0.001
Cohen's d	1.26 (large)
95% CI (Mean)	[3.74, 4.28]
Holm-Bonferroni $p_{\text{corrected}}$	< 0.001
Significance ($\alpha = 0.05$)	Significant
Wilcoxon Signed-Rank	$W = 31.0$
p -value (raw)	< 0.001
Effect size r	-0.781 (large)
Median Diff. from Neutral	1.00
Holm-Bonferroni $p_{\text{corrected}}$	< 0.001
Significance ($\alpha = 0.05$)	Significant

Qualitative feedback further illustrated how participants used featured questions (CuQAs) to form judgements about the breadth and limits of the KG. Several participants commented on the perceived range and usefulness of the information surfaced, noting, for example, that specific questions yielded "particularly useful" results or, that the questions were "very intuitive" albeit in need of expansion. Others explicitly reflected on the apparent scope of the KG, observing that a limited set of people-related questions suggested narrower coverage, while place-related questions conveyed a broader geographical range. These reflections align closely with the large quantitative effects observed for scope uncertainty, indicating that featured questions (CuQAs) functioned as a way of helping users infer what types of content the KG did, and did not, contain.

5.6.4.5 Ontology Opacity Analysis: Post Use of Featured Questions (CuQAs)

The ontology opacity construct also demonstrated good internal reliability (Cronbach's $\alpha = 0.75$) suggesting that the composite consistently captured perceived approachability and ease of exploration. Similar to scope uncertainty, Shapiro-Wilk indicated that the data significantly deviated from normality ($p = 0.0093$) but once again, both parametric and non-parametric results were reported for completeness.

Ontology opacity was analysed using the same neutral-point comparison approach (Table 5.9). The one-sample t -test showed that the post-interaction mean (mean = 4.11, 95% CI [3.87, 4.36]) was significantly above the neutral midpoint ($t(34) = 9.36$, $p \leq 0.001$), with a very large effect size (Cohen's $d = 1.58$). The Wilcoxon signed-rank test likewise indicated a median of 4.33 and a large effect size ($r = -0.831$, $p \leq 0.001$).

These results remained statistically significant after Holm-Bonferroni correction

($p_{\text{corrected}} \leq 0.001$). Together, they indicate that featured questions (CuQAs) were associated with substantial positive perceptions of the KG’s structure when compared to neutral, making the KG feel more approachable and easier to explore without requiring detailed prior understanding of its underlying structure.

Table 5.9: Ontology Opacity: Post-Use of Featured Questions (CuQAs) vs. Neutral

Metric	Value
Sample size (n)	35
Mean (SD)	4.11 (0.70)
Median	4.33
Neutral Reference	3.00
Cronbach’s α	0.75 (good)
Item-total correlations (r range)	0.67 – 0.92
One-sample t -test	$t(34) = 9.36$
p -value (raw)	< 0.001
Cohen’s d	1.58 (large)
95% CI (Mean)	[3.87, 4.36]
Holm-Bonferroni $p_{\text{corrected}}$	< 0.001
Significance ($\alpha = 0.05$)	Significant
Wilcoxon Signed-Rank	$W = 14.0$
p -value (raw)	< 0.001
Effect size r	-0.831 (large)
Median Diff. from Neutral	1.33
Holm-Bonferroni $p_{\text{corrected}}$	< 0.001
Significance ($\alpha = 0.05$)	Significant

5.6.5 Summary of Key Findings

From the results of the evaluation, **H1_UE2** "CuQAs are perceived to address the observable barriers that characterise the IEP for lay users unfamiliar with the VRTI KG" is partially supported.

The extent to which can be understood by looking more specifically at the associated research question: **RQ1_UE2**: "To what extent do CuQAs address perceived scope uncertainty, perceived ontology opacity, and improve perceived orientation for lay users unfamiliar with the VRTI KG?". Across the three IEP barriers and emergent outcome, distinct patterns of impact were observed. For orientation, no statistically meaningful change was detected between pre- and post-use of featured questions (CuQAs). This absence of effect is plausibly explained by the high baseline levels of perceived orientation, which limited the scope for measurable improvement and suggests the presence of a ceiling effect driven by both existing interface affordances and participant domain familiarity. In other words, a ceiling in perceived orientation does not imply absence of underlying epistemic barriers, but rather indicates that alternative supports had already elevated users’ sense of entry confidence.

This is in contrast to both scope uncertainty and ontology opacity where large, statistically robust improvements were exhibited when compared against the neutral reference

point. Participants consistently reported a clearer understanding of the range and types of information contained within the KG, as well as an increased sense that the KG was approachable and navigable without specialised background knowledge.

Overall, while CuQAs did not measurably enhance orientation in a cohort that already felt relatively confident in knowing where to begin, they demonstrably addressed two fundamental and theoretically grounded barriers of the IEP when compared against neutral. The magnitude and consistency of the observed effects for scope uncertainty and ontology opacity provide strong empirical support for the role of CuQAs as an effective mechanism for facilitating early-stage exploration of complex KGs in the humanities domain.

5.7 LLM Experiment 1: TMv1 Stage 1, LLM-supported CQ Template Derivation

LLM Experiment 1 investigated how well LLMs can carry out the derivation of CQ templates by evaluating their suitability for Stage 1 of the example instantiation of Tús Maith, TMv1 (see Figure 1.2). This section makes use of the content contained within, and expands upon, the paper published by the author of the thesis [44].

While choosing LLMs as the method of instantiation for the Tús Maith framework appeared to be a promising approach to test its viability (as outlined by the reasons presented in Section 5.16 above), it was imperative to assess how they fared in practice i.e., the extent to which they can generate clear, relevant, and answerable CQ templates when provided with a KG’s ontology. One of the primary motivating factors in choosing LLMs was their potential ability reduce the cognitive load of the domain expert(s) tasked with creating CuQAs, that LLMs would reduce the need for domain experts to recall the entirety of the contents of the KG in order to come up with CQ templates. While cognitive load itself was not examined as part of this experiment, a manifestation of this benefit would be if the LLM could surface plausible CQs that had not been thought of by a domain expert alone, i.e. increase coverage of the set of CQ templates. Such thoughts and considerations led to the hypotheses outlined in the next section.

5.7.1 Hypotheses

The following hypotheses were examined in the experiment:

H1_EX1: LLMs at their default settings (temperature=1, top_p=1) can generate relevant, clear, and answerable CQ templates suitable for Tús Maith when provided with information about the VRTI KG’s ontology. Suitable for Tús Maith here means that the CQs generated, when instantiated, could act as exploratory starting points for non-domain familiar lay users.

H2_EX1: LLMs can increase the coverage of a set of CQ templates for the VRTI KG (as judged by domain experts relative to an expert-authored GT set).

5.7.2 Experimental Setup

The Experimental Setup section details how the evaluation was designed and what it consisted of, how the data was prepared, and what exactly was being examined. The LLMs chosen for this evaluation were Google’s Gemini 2.0 Flash [54] and Open AI’s GPT-4o [52]. These were chosen as they were widely deployed, instruction-following LLMs available at the time of the experiment. Their use enabled the evaluation to focus on the feasibility and limitations of the proposed instantiation under practical conditions, rather than on optimising performance for a particular model. Since the task involved generating linguistically well-formed, abstract question templates from provided ontology terms rather than retrieving factual knowledge, the evaluation emphasised generative adequacy over model-specific knowledge coverage.

5.7.2.1 Evaluation Strategy

This subsection outlines the two-step evaluation process undertaken as part of this experiment. At a high level, following sequential explanatory design [112], the number of question sets generated by the LLMs was first reduced from 12 sets per LLM (arising out of 12 prompt variants explained in detail below in Section 5.7.2.3) to 1 set per LLM (2 sets total) by keeping the most semantically similar set to a GT question set (discussed in the next section) using SentenceBERT [42] (explained in detail below). In utilising SentenceBERT, the aim was not to identify the "best" questions in absolute terms, but to select candidate sets that were semantically anchored to expert-authored coverage while still allowing for variation and novelty. Those 2 question sets were then evaluated via a questionnaire across three dimensions (relevance, clarity, and answerability) using both quantitative and qualitative measures by two domain experts (DE1, and a new domain expert, also intimately familiar with the VRTI KG who shall hereafter be referred to as DE3) (H1_EX1). Further detail of this can be found in Section 5.7.3.2. The domain experts (DE1 and DE3) were also asked to evaluate whether the generated set of CQs contained questions that, in their opinion, could be added to the GT question set (H2_EX1). While step two, the domain expert evaluation, included both Likert-scale ratings and open-ended responses, its primary role within the mixed-methods design is explanatory. The structured ratings provided a quantitative indication of expert perceptions, while the accompanying qualitative feedback was used to interpret, contextualise, and explain these ratings in contrast to the earlier computational reduction step which determined the specific artefacts subjected to expert review. Furthermore, it was deemed infeasible to give all 24 generated LLM question sets to the domain experts to evaluate which is why the question sets were reduced in the first place. This aligned with the experiment’s goal of evaluating LLMs in their capacity to generate template CQs as exploration starting points rather than strictly a prompt fine-tuning exercise. Crucially, the experiment did not aim to rank LLMs, but to evaluate the feasibility and limitations of using LLM-supported generation as an instantiation mechanism for Stage 1 of Tús Maith.

The experimental design was inspired by the methodology presented in [99], which evaluated six LLMs across five RDF-based ontologies using zero- and few-shot prompting.

While the LLM experiment here for the thesis involved only two LLMs, a greater variety of prompts was used (12 distinct prompts as opposed to 6), and crucially, domain experts evaluated the output. The methodology was further informed by the work of Antia et al. [140], who explored the automatic generation of CQs using a Text-to-Text Transformer model [141] and emphasised expert-based assessment of CQ quality. While their approach does not rely on LLMs, it provides a methodological precedent for grounding automated CQ generation in domain-expert validation, which this study adopts and extends in the context of LLM-driven systems. The full results are available in the thesis GitHub repository¹ in the LLM Experiment 1 folder.

Step 1 - SentenceBERT Evaluation and GT Acquisition As there was no existing benchmark CQ set for the VRTI KG, to establish a minimal GT, DE1 was asked to create a set of CQs that they considered captured the breadth of the KG (Table 5.10).

Table 5.10: Domain Expert Authored CQ Templates

ID	Question
1	When was a person born?
2	Where was a person born?
3	When did a person die?
4	Where did a person die?
5	What century was a person active in?
6	Which people have this surname?
7	What type of place is this?
8	What period of time/era does this placename belong to?
9	Does this person have a Dictionary of Irish Biography entry?
10	What county is a parish/townland/barony located in?

Once obtained, the LLM generated CQ templates were first evaluated for semantic similarity to the expert-authored ones with SentenceBERT using the all-mpnet-base-v2 sentence-transformer model [120] and compared pairwise using cosine similarity.

Let $S \in \mathbb{R}^{m \times n}$ represent the similarity matrix, where m is the number of generated questions and n the number of expert-authored ones. Each element s_{ij} corresponds to the similarity between the i -th generated question and the j -th expert question.

For each generated question, the maximum similarity to any expert question was calculated:

$$\max_i = \max_{1 \leq j \leq n} s_{ij}$$

The average of these maximum similarity scores across all generated questions was then computed as:

$$\overline{\max} = \frac{1}{m} \sum_{i=1}^m \max_i$$

To assess how many generated questions achieved a similarity above a chosen threshold τ ,

¹https://github.com/mcnamacl/T-s_Maith_Thesis_Repository

the following is calculated:

$$c = \sum_{i=1}^m \mathbb{I}(\max_i \geq \tau)$$

where $\mathbb{I}$ is the indicator function. The percentage of generated questions exceeding the threshold is given by:

$$p = \left(\frac{c}{m} \right) \times 100$$

In this experiment, a threshold of $\tau = 0.6$ was selected, following precedent established in [99] evaluating LLM-generated CQs using SentenceBERT. This value reflects a moderate level of semantic alignment, capturing cases where generated questions are meaningfully similar to expert-authored ones while avoiding overly lenient matches.

Step 2 - Expert Evaluation While semantic similarity provides an automated means of filtering and ranking generated questions, it does not assess important qualitative aspects such as interpretability, alignment with the KG’s intended usage, or answerability using only the information found within the KG. To address this, a human expert evaluation was conducted on a subset of the generated questions. Prior to evaluating the generated questions with DE1 and DE3, an ethics application was sent in and approved by the SCSS Research Ethics Committee. The full ethics application can be found in Appendix A.4.3. DE1 and DE3, were recruited via email (Appendix A.1.2) and asked to evaluate a sample of the LLM-generated questions. This sample comprised of one set of CQ templates per LLM, the sets of questions that achieved the highest average semantic similarity to the expert-authored GT questions for each LLM out of the 12 prompt configurations described in Section 5.7.2.3. Following [140], each question was evaluated independently by both experts using a 5-point Likert scale [130] across three dimensions:

1. **Relevance:** How well does the question align with the purpose of a CQ in this context? (1 = Not relevant at all, 5 = Highly relevant)
2. **Clarity:** Is the question phrased in a clear, unambiguous, and grammatically correct manner? (1 = Very unclear or confusing, 5 = Very clear)
3. **Answerability:** Can the question be reasonably answered using information contained within the VRTI KG? (1 = Not answerable at all, 5 = Fully answerable)

These three dimensions were selected to reflect the practical requirements of CQ templates within the context of the Tús Maith framework – question templates that highlight the scope and potential of a KG that can later be filled in with specific entity information. *Relevance* considers whether the question aligns well with the definition of a CQ template. *Clarity* addresses interpretability, particularly for non-technical users. *Answerability* considers whether the question can reasonably be answered using only the information currently present in the KG.

The experts were instructed to evaluate each question as a CQ template, where a CQ was described in the questionnaire as: "For the purposes of Tús Maith, a competency question is a natural-language template question that expresses what the knowledge graph should be able to answer. These questions give users a sense of the breadth, scope, and potential of the knowledge graph before they begin exploring it. The intention is that these questions can later be filled in with popular entity information e.g. a CQ template might look like: Where was a person born? This question can later be filled in with popular person entity information such as: Where was Oscar Wilde born? You are being asked to evaluate the template questions.". The link to the full questionnaire can be found in Appendix A.2.1. In addition to scoring individual questions, the experts were also invited to reflect on whether any questions present in the generated set but not in the original GT set (Table 5.10) could, in hindsight, be considered CQ template candidates. This qualitative feedback was intended to identify whether the LLM could highlight potential blind spots in the expert-authored GT CQs. Expert scores were aggregated to provide average scores per dimension across each LLM-generated set. This enabled comparison not only between individual questions but also between the two different LLMs used. This evaluation stage was essential in order to identify whether either/both/or neither LLM could produce not only semantically aligned questions with the GT CQs, but ones that were perceived by experts as meaningful, understandable, and practically useful in the context of supporting exploration within a complex KG. While DE1 constructed the GT set, both experts evaluated the generated questions independently. The GT set was used both for the computational filtering step and as a reference point for assessing coverage, rather than as a prescriptive evaluation rubric.

5.7.2.2 Data Preparation

The VRTI KG integrates multiple ontologies, including CIDOC CRM [12], GeoSPARQL [13], and a custom VRTI ontology [14], making it significantly more complex than single-schema KGs. CIDOC CRM, in particular, is a formal ontology designed to represent cultural heritage information in rich detail, featuring a dense network of abstract classes and properties that model events, entities, and relationships. The resulting heterogeneity obtained through the use of multiple, often complex ontologies poses additional challenges for CQ template generation, as many terms are either highly abstract or domain-specific (e.g. `P81b_begin_of_the_end`). To explore whether different input scopes could yield more interpretable and relevant questions, three variations of input term sets were tested across prompt runs: the full set of classes and properties (48 classes, 198 properties), a manually reduced subset (referred to as "Manual Subset") of 18 classes and 35 properties, and a third version of the subset that included CIDOC CRM relationship information to the subset of classes and properties as prompt variables. All input variations can be found in the thesis GitHub ² in the LLM Experiment 1 folder. The Manual Subset was constructed by the author of this thesis to retain the most semantically expressive and user-relevant terms while excluding overly technical entries, such as `virtrdf#qmf01blank0fShortTpl`,

²https://github.com/mcnamacl/T-s_Maith_Thesis_Repository

which are poorly suited to natural language articulation. The decision to extract classes and properties from instance data rather than from the full ontology was intentional; not all ontology terms are currently used in the KG, and this work aims to support exploratory access to available content, rather than exhaustively covering the ontology itself.

5.7.2.3 Prompt Design and Experimental Setup

Twelve separate runs were conducted for each LLM, varying by prompt framing, shot type (zero-shot or few-shot), and input scope as seen in Table 5.11. Input scope options are discussed in the previous section.

1. **Prompt 1:** Simply instructs the LLM to generate questions from provided classes and properties.
2. **Prompt 2:** Includes the assigning of a role to the LLM (historical KG exploration assistant) to guide its generation strategy.

Table 5.11: The run numbers corresponding to each combination of prompt type, shot type, and input scope.

Prompt Type Shot Type	All Classes & Properties	Manual Subset	Manual Subset + CIDOC CRM Relationship Information
Prompt 1 Zero-shot	Run 1	Run 2	Run 3
Prompt 1 Few-shot	Run 4	Run 5	Run 6
Prompt 2 Zero-shot	Run 7	Run 8	Run 9
Prompt 2 Few-shot	Run 10	Run 11	Run 12

Listing 1 in the Appendix Section A.5.1 shows the base prompt used for all LLM runs, with conditional sections annotated according to experimental configurations (e.g., +Role, +Examples).

The number of CQ templates requested from each LLM run was fixed at 20. This value was selected to align with the conceptual structure of the VRTI KG, which at the time of the experiment comprised 18 semantically meaningful classes within the manually reduced ontology subset. Rather than enforcing a one-to-one correspondence between classes and questions, a slightly higher target was chosen to allow for both class-level coverage and the emergence of cross-class or relational questions that span multiple conceptual categories (e.g., person-place). This design provided limited redundancy, enabling expert reviewers to assess not only whether individual classes were represented, but also whether the LLM surfaced novel or integrative CQ templates beyond a minimal coverage baseline. As such, the choice of 20 questions reflects a coverage-oriented sampling strategy rather than an

attempt to optimise model performance or exhaustively enumerate the space of possible CQs.

5.7.3 Results

5.7.3.1 SentenceBERT Similarity Evaluation

Firstly, the similarity between the 20 questions generated in each run (examples of which can be found in Table 5.14) and the set of 10 domain expert-created minimal GT questions (Table 5.10) was obtained using SentenceBERT. The metric used was maximum similarity per question, with average values reported per run in Table 5.12. Across all runs, the highest-performing runs were Run 6 (the few-shot prompt with the manual subset of the classes, properties and CIDOC CRM relationship information) for GPT-4o, and Run 11 (the few-shot prompt with a role assigned and just the manual subset of classes and properties) for Gemini 2.0 Flash. The best run for GPT achieved an average max similarity of 0.418 (20% $\geq$ 0.6), while the best for Gemini achieved 0.480 (25% $\geq$ 0.6), indicating a closer lexical and semantic match to the expert-authored questions.

Table 5.12: Average semantic similarity score (as computed using SentenceBERT) per run, and the number of generated questions per run that met or exceeded the 0.6 similarity threshold.

Run	GPT Avg. Sim	GPT $\geq$ 0.6	Gemini Avg. Sim	Gemini $\geq$ 0.6
1	0.332	0/20	0.382	2/20
2	0.306	1/20	0.378	2/20
3	0.415	4/20	0.396	3/20
4	0.367	2/20	0.330	0/20
5	0.390	2/20	0.366	2/20
6	0.418	4/20	0.422	3/20
7	0.354	2/20	0.393	2/20
8	0.334	0/20	0.430	3/20
9	0.403	2/20	0.380	4/20
10	0.345	2/20	0.378	1/20
11	0.389	3/20	0.480	5/20
12	0.337	1/20	0.461	4/20

Notably, runs with access to CIDOC CRM relationship data (runs 3, 6, 9, and 12) did not consistently outperform others, and runs provided with the full set of classes and properties tended to underperform. This may suggest that narrowing the input scope to a manually filtered subset, without CIDOC CRM relationship data, helps steer LLM generation toward more meaningful or semantically focused questions.

5.7.3.2 Domain Expert Evaluation

For step two, the domain expert evaluation, question sets generated by runs 6 (GPT) and 11 (Gemini), the highest-scoring runs in Table 5.12, were used. Two domain experts, DE1 and DE3, independently expressed their judgment on these question sets, rating each individual question via an online questionnaire (the full questionnaire can be found in

Appendix A.2.1) on relevance, clarity, and answerability using a 1-5 Likert scale. They were also asked whether any questions in the generated sets could be added to the GT set, to select which question set they preferred overall, and to provide free-text comments. The experts were not told which question set was generated by which LLM. The average scores and quantitative analysis of these scores across both experts are summarised in Table 5.13.

Table 5.13: Average (Avg) and standard deviation (Std Dev.) of the domain expert (DE1 and DE3) ratings for the two question sets (Gemini and GPT) across three evaluation dimensions: relevance, clarity, and answerability; the inter-rater reliability for each dimension, measured using intraclass correlation coefficients (ICC 3,k) [118], and the confidence interval (CI) for each ICC.

Metric	Model	DE1 Avg., SD	DE3 Avg., SD	OverallAvg.	ICC(3,k)	95% CI
Relevance	Gemini	2.55, 1.669	3.2, 1.542	2.875	0.144	[-1.16, 0.66]
Relevance	GPT	3.05, 2.038	3.4, 1.273	3.225	-0.015	[-1.56, 0.60]
Clarity	Gemini	2.95, 1.701	3.75, 1.372	3.35	0.539	[-0.16, 0.82]
Clarity	GPT	3.25, 1.943	3.95, 1.099	3.6	0.839	[0.59, 0.94]
Answerability	Gemini	2.35, 2.159	3.85, 1.268	3.1	0.671	[0.17, 0.87]
Answerability	GPT	1.95, 2.114	3.15, 0.988	2.55	0.839	[0.59, 0.94]

While the GPT-generated question set scored slightly higher in average relevance and clarity, the Gemini-generated question set outperformed it in answerability, based on mean ratings across both experts. Inter-rater reliability varied significantly across dimensions and LLMs. For relevance, both models showed extremely low agreement between raters, with intraclass correlation coefficients (ICCs) close to zero or even negative (GPT ICC3k = -0.015, Gemini ICC3k = 0.144), indicating substantial subjectivity in how relevance was interpreted. This is further reflected in the high standard deviations (up to 2.038 for GPT), suggesting that domain experts varied widely in their opinion on how well the questions align with the purpose of a CQ in the context of Tús Maith (relevance). In contrast, GPT achieved strong inter-rater agreement for both clarity (ICC3k = 0.839) and answerability (ICC3k = 0.839), accompanied by comparatively lower standard deviations, indicating more consistent scoring and tighter consensus among experts. Gemini exhibited moderate reliability (clarity ICC3k = 0.539; answerability ICC3k = 0.671) but showed greater variability in scores, especially for answerability, where standard deviations reached 2.159 for DE1. These combined metrics – ICCs capturing rating agreement between the domain experts, the average rating given by them to the generated question sets across the three criteria, and standard deviations indicating score dispersion – highlight the nuanced trade-offs between average performance and consistency of expert assessments, which is critical in qualitative evaluations with subjective criteria.

Expert ratings reflected common patterns: questions rated poorly typically contained ambiguous wording or technical jargon (e.g. "What is the preferred identifier of a person?") that obscured their intent and reduced accessibility for non-technical users. In contrast, highly rated questions were clear, focused on meaningful historical information, and aligned well with the VRTI KG, making them both understandable and answerable (e.g. "When was a person born?"). However, these findings also highlight the need to analyse questions

comprehensively across all three dimensions rather than focusing on any single criterion. For example, some questions, such as "*What is the relationship between a group and a person?*", were rated highly for relevance and clarity by both domain experts but scored lower on answerability due to it referring to information not explicitly contained within the VRTI KG. This underscores that a balance across all dimensions is essential for a question to function well as a CQ Template, generated from Stage 1. Table 5.14 presents paired examples of questions across all three evaluation dimensions (Relevance, Clarity, Answerability), showing examples of both the lowest-rated ($\leq 2/5$) and highest-rated ($\geq 4/5$) questions where both domain experts agreed in their assessments.

Table 5.14: Comparison of question quality between Gemini and GPT models across the three evaluation criteria (Relevance, Clarity, Answerability), showing representative examples rated highly and poorly by both domain experts.

Metric	Rating	Gemini Example	GPT Example
Relevance	Poorly Rated	<i>What type of human-made thing is this?</i>	<i>Who carried out the attribute assignment?</i>
	Highly Rated	<i>When was a person born?</i>	<i>What role did a person have?</i>
Clarity	Poorly Rated	<i>What is the preferred identifier of a person?</i>	<i>Who is identified by this identifier?</i>
	Highly Rated	<i>What is the hectare size of a place?</i>	<i>What is the century associated with an event?</i>
Answerability	Poorly Rated	<i>What is depicted by a physical thing?</i>	<i>What elements compose this object?</i>
	Highly Rated	<i>What place is a birth associated with?</i>	<i>What is the gender of a person?</i>

Expert comments offer important qualitative insights. While both models generated syntactically valid questions, they were often overly complex or used domain-specific jargon unsuited to non-technical users, for example, "*What type of attribute assignment was carried out?*". One expert observed that Gemini’s questions, although still technical, were easier to interpret and more closely matched the structure and content of the VRTI KG. This likely contributed to both experts ultimately preferring the Gemini-generated set. A notable outcome of the evaluation is that each expert identified at least one Gemini question that could be added to the GT CQ set and one expert did so for GPT. An example of a question a domain expert found that could be added is "*What title does a person have?*". This indicates that LLM-generated questions, despite the current limitations, can reveal potential gaps in manually constructed question templates.

5.7.4 Comparative Insights

The comparison between SentenceBERT similarity and expert judgment reveals both alignment and divergence. GPT, despite slightly lower SentenceBERT similarity than Gemini, was rated more relevant and clear but less answerable. This suggests that it may generate

more ambitious or abstract questions. Conversely, Gemini’s higher similarity and answerability scores indicate that it may produce questions closer to the current graph content and better suited to near-term user exploration needs. An important confounding factor may be KG class coverage. 7 of 10 (70%) expert GT questions were person entity related, whereas only 9 of 20 (45%) Gemini questions and 10 of 20 (50%) GPT questions in their respective best runs referred to people.

5.7.5 Summary of Key Findings

H1_EX1: "LLMs at their default settings (temperature=1, top_p=1) can generate relevant, clear, and answerable CQ templates suitable for Tús Maith when provided with information about the VRTI KG’s ontology." is partially supported: the results highlight that while LLMs can generate semantically rich and partially relevant CQs, there is a trade-off between semantic similarity, readability, and groundedness in the graph. Narrowing input scope and including few-shot examples improved performance, and SentenceBERT evaluation aligned in part with expert assessments. However, inter-rater reliability metrics revealed variation in how experts interpreted key criteria like relevance, highlighting the need for clearer evaluation rubrics for domain experts to follow. While some subjectivity is inevitable, even among domain experts working in the same area provided with a clearer rubric, relevance still remains a valuable criterion as it is one that situates the question within the specific context of CQ templates, in ways that clarity and answerability alone cannot capture. Additionally, while the clarity criterion, like the others, implicitly includes multiple subcomponents (e.g., grammar, ambiguity, phrasing), no formal rubric was used to disaggregate or define these dimensions. The evaluation instead reflects common humanities practice, where expert interpretation and contextual judgment are often more appropriate than rigid scoring frameworks [142]. Nonetheless, the extremely low inter-rater agreement for relevance seen here prompted the breaking up of the criterion into three more specific questions for User Evaluation 3 (Section 5.9).

H2_EX1: "LLMs can increase the coverage of a set of CQs for the VRTI KG." is supported: both domain experts identified a CQ template generated by an LLM that they would add to the existing expert-authored set of CQs. This evidence indicates that LLMs might act as a cognitive aid to domain experts in Stage 1 of TMv1 generating CQ templates aligning with the core design principle of LLMs acting as scaffolds, not substitutes for domain experts.

5.8 LLM Experiment 2: TMv1 Stage 2, LLM-supported CQ Template Instantiation

LLM Experiment 2 examined how well LLMs can act in Stage 2 of TMv1, CQ Template Instantiation (see Figure 1.2). Similar to Section 5.7, the content of this section makes use of and expands upon published work by the author of this thesis [45].

5.8.1 Hypothesis

The hypothesis related to LLM Experiment 2 is the following:

H1_EX2: LLMs at their default settings (temperature=1, top_p=1) can batch instantiate CQ templates to produce filled-in questions and answers using information about multiple entities.

Satisfying this hypothesis was a requirement for an LLM to be included in Stage 2 of TMv1.

The need for explainability and post-generation verifiability in a curator-facing CuQA LLM-assisted system (TMv1) motivated the requirement for structured JSON output and explicit provenance. Reporting the source entity URI established a minimal, machine-checkable link between each filled-in question and answer and the entity used to instantiate it, enabling automated validation and human curator oversight. This aligns with established work in explainable and trustworthy AI, which emphasises that model outputs should be accompanied by inspectable evidence rather than treated as opaque predictions [143, 144]. A stronger provenance condition, in which the LLM also reports the specific key-value evidence used for slot filling, prioritises human interpretability but may have imposed additional generation overhead; this trade-off was therefore examined empirically [145].

In light of this, the following subsection breaks the hypothesis into measurable research questions to assess the extent to which LLMs can satisfy H1_EX2.

5.8.2 Problem Definition and Research Question

This section formally defines the computational task under investigation and presents the specific research questions that guided this experiment. In order to differentiate these research questions from the thesis’ research question, they have been given the acronym RQX_Exp2, where X denotes the question’s number.

5.8.2.1 Task Definition

The core task evaluated in this experiment is the **LLM-supported instantiation of CQ templates across multiple entities**. Formally, the task is defined as follows:

- **Input:**

- A set of CQ templates $\mathcal{T} = \{T_1, T_2, \dots, T_m\}$
- A structured JSON object containing all information about multiple entities $\mathcal{E} = \{e_1, e_2, \dots, e_n\}$, where each entity e_i is identified by a unique URI and associated key-value pairs

- **Output:** For each entity-template pair where the template is applicable (i.e., the entity contains the necessary properties to answer the question), a natural language question-answer pair (Q_{ij}, A_{ij}) where:

- Q_{ij} is the instantiated question created by filling template T_j with information from entity e_i

- A_{ij} is the corresponding answer derived from the information about e_i
- **Constraints:** The output must adhere to strict constraints of:
 - **Entity Fidelity:** Each answer A_{ij} must be verifiably grounded in the provided information about entity e_i .
 - **Slot Coverage:** All variable slots in each template T_j must be correctly filled by information from the relevant entity e_i .
 - **Prompt Compliance:** The LLM’s output must strictly follow the specified format (e.g., including provenance information as required).

This batch processing task requires the LLM to manage multiple template question-answer pair instantiations while maintaining consistency and accuracy across outputs.

In order to be able to thoroughly assess the LLM output beyond just slot level metrics, using metrics like semantic similarity via SentenceBERT for example, there needed to be a GT with which to compare the generated output against. A rule-based, silver-standard, programmatic GT was therefore constructed to provide a baseline with which to compare the generated instantiated question-answer pairs against. This is discussed in more detail in Section 5.8.3.2. The dependency on a GT gave rise to the question of - how can the generated output be validated in the absence of a GT? In other words, when the end-to-end system TMv1 is used in practice, there is more than likely not going to be a GT that can be used to compare LLM generated output against. In answering this question, the idea of to create a JSON validation proxy using both the structured JSON input and using the URI provenance information returned by the LLM for corresponding to the entity information they used to instantiate the question-answer pair was conceptualised. This is discussed in more detail in Section 5.8.3.6.

5.8.2.2 Research Questions

To systematically evaluate the performance of LLMs on this batch instantiation task, the investigation was framed around the following primary research question:

RQ1_Exp2: How effectively can LLMs perform batch instantiation of CQ templates and their answers across multiple entities while maintaining accuracy, fidelity, and robustness? This evaluation spans four key dimensions: (1) accuracy and fidelity of slot filling, (2) completeness of question coverage, (3) robustness against hallucination and prompt non-compliance, and (4) efficiency in runtime.

The primary RQ1_Exp2 is decomposed into the following specific sub-questions:

RQ1.1_Exp2 (Provenance Impact): How does requiring different levels of provenance evidence (basic: entity URI only vs. full: URI + explicit key/value evidence) affect LLM performance across accuracy metrics (slot-level exact match, entity-level correctness), semantic similarity measures (SBERT, BERTScore), hallucination rates, and efficiency (run times)?

RQ1.2_Exp2 (Validation Proxy): To what extent can automated JSON validation against source entity data serve as a practical proxy for a programmatically created Ground Truth (GT) when evaluating batch instantiation performance?

5.8.3 Experiment Methodology

The LLM’s performance on the template-filling question-answer pair task was evaluated using a comprehensive pipeline that balanced ecological validity with rigorous evaluation standards. The methodology presented here covers data and entity selection, GT establishment, LLM inference, and multi-dimensional evaluation.

More specifically, this methodology adopts a task-grounded, system-oriented evaluation perspective, assessing LLMs not as standalone instantiation models but as functional components within a CuQA curation system. Design choices prioritise ecological validity by grounding entity selection in observed user behaviour rather than synthetic benchmarks, aligning with human-centred evaluation practices that emphasise realism and contextual fidelity in system assessment [146, 147]. Controlled benchmarking is ensured through deterministic GT construction and calibrated semantic evaluation metrics, following evidence-centred and benchmark design principles in NLP that require explicit alignment between task design, measurement, and interpretive claims [148, 42, 43]. KGQA evaluation literature and entity-ranking studies further motivated the choice of metrics and entity selection heuristics including the use of exact and lexical matching, semantic similarity, and entity ranking [149]. Accordingly, the evaluation combines token-level (BERTScore [43]) and sentence-level (SBERT [42]) measures with strict slot checks to capture both lexical fidelity and semantic equivalence. Finally, a deployment-oriented validation layer is introduced via a lightweight, GT-agnostic JSON proxy inspired by weak supervision and schema-based validation paradigms, enabling scalable quality control under realistic operational constraints [150, 151, 152].

The full code, datasets created, and reproducibility procedure for the data preparation and evaluation can be found in this GitHub repository³ in the LLM Experiment 2 folder.

5.8.3.1 Data and Entity Selection

At the time that the experiment was taking place, the VRTI Explorer [37] had been launched publicly in June 2025 and had been live for seven weeks. To leverage the real-world applicability of the experiment, the top 20 search terms from VRTI Explorer Google analytics [153] were retrieved and served as the basis with which to identify entities for the experiment. Similarly, nine of the ten historian-curated CQ templates generated as part of Section 5.7 were used as the CQ templates to evaluate the LLMs (see Table 5.15). The CQ template "Which people have this surname?" was removed as it required the reverse of the logic operationalised as part of this experiment i.e. it asked for a list of persons rather than starting with a list of persons and finding their properties.

³https://github.com/mcnamacl/T-s_Maith_Thesis_Repository

Retrieving and normalising search terms As mentioned, the top 20 search terms were retrieved via the Google Analytics API [153]. As several terms mapped to the same entities (e.g., "James" and "James "), a normalisation step was required to deduplicate the search terms. After this preprocessing, only 19 unique search terms remained. To address this, an overfetch factor of 5 was implemented: the API returns the top 20×5 terms, which were iteratively normalised until 20 unique search terms were obtained. Search terms that did not resolve to any entity in the KG (e.g., "james brice robinson") were excluded and replaced with the next highest-ranked entity. An integrated ASK SPARQL query (which can be found in Appendix A.8.2) ensured that each selected search term corresponded to at least one KG entity.

Entity ranking Due to search term ambiguity, entity ranking was necessary [149]. For example, "James" mapped to 700+ entities within the KG. An entity ranking algorithm therefore was developed selecting a maximum of 10 top entities per search term (5 persons, 5 places) based on:

- **Text relevance (*matchScore*)**: a rule-based similarity score prioritising exact, token, prefix, and substring matches with the search term.
- **KG richness (*coverageScore*)**: a weighted sum of normalised slot coverage (0.5), external links (0.25), and graph connectedness (0.25).

The final score is calculated as: $\text{finalScore} = \text{matchScore} \times (0.5 + 0.5 \times \text{coverageScore})$, which ensures text relevance is primary while still promoting entities with richer descriptions. The full entity ranking algorithm can be found in the Appendix A.6.

Stratification From the ranked list, entities were stratified into three equal tiers according to their finalScore: High (≥ 1.0102), Medium (0.7000-1.0102), and Low (< 0.7000). Tiers were defined using quantiles from the sorted scores: Low ($< \text{low_threshold}$, bottom 34%), Medium (low_threshold to high_threshold , middle 33%), and High ($\geq \text{high_threshold}$, top 33%), where low_threshold was the 0.34 quantile (34th percentile) and high_threshold was the 0.67 quantile (67th percentile). The final evaluation pool consisted of 17 batches, each containing 10 entities (5 persons, 5 places), balanced across tiers ensuring that no one batch would be more or less rich for the LLM to work with than another. An 18th incomplete batch was rejected due to there not being enough entities left over to form a complete batch. All batches used can be found in the GitHub repository.

5.8.3.2 Ground Truth (GT) Construction

For each batch, SPARQL queries (see Appendix A.8.4) retrieved entity information serialised to JSON, serving as input for a programmatic GT construction, LLM source data, and validation reference. Table 5.15 shows the template-to-JSON field mapping.

For each entity and CQ template pair, answers were extracted from the source JSON and normalised using deterministic rules:

Table 5.15: CQ Templates and Corresponding Answer Sources from Source JSON

Template Question	Answer Source
"When was {Person,Name} born?"	DisplayBirthDate
"Where was {Person,Name} born?"	BirthPlaceLabel
"When did {Person,Name} die?"	DisplayDeathDate
"Where did {Person,Name} die?"	DeathPlaceLabel
"What century was {Person,Name} active in?"	FloruitLowerYear & FloruitUpperYear
"What type of place is {Place,Name}?"	CRM_Type
"What period of time/era does {Place,Name} belong to?"	URI Parsing
"Does {Person,Name} have a Dictionary of Irish Biography entry?"	DIB_Page
"What county is {Place,Name} located in?"	FallsWithinPlace

- Dates were converted to ISO format (YYYY-MM-DD).
- Floruit periods (e.g., FloruitLowerYear: 1750, FloruitUpperYear: 1780) were generalised to centuries (e.g., "18th century"). In the event that the floruit years crossed centuries (e.g., FloruitLowerYear: 1790, FloruitUpperYear: 1830), both centuries were recorded as "18th century-19th century".
- Place types from CRM [12] class IRIs (e.g., "<https://ont.virtualtreasury.ie/ontology#PresentDayTownland>") were converted to natural language labels (e.g., "Present Day Townland").
- Geographic hierarchies (e.g., a townland's county) were resolved through label parsing and source URL analysis.

This method provided a deterministic, reproducible, and objective baseline for evaluating LLM performance on the task of KG-grounded question answering. Furthermore, it served as a rule-based design alternative to consider. A key limitation of the rule-based instantiation is that the programmatic rules, while robust, can be brittle: any change in the ontology, addition of new templates, or new entity types requires manual rule adjustments, demanding technical expertise, which contradicts TMv1's goal of supporting non-technical domain experts. In conjunction with this, the approach may occasionally lack the nuance of a human (or even an LLM) instantiator (e.g., omitting hierarchical detail in place names: London, England vs. London for example). Table 5.16 provides an overview of the trade-offs between the instantiation methods that ultimately led to performance of the experiment leveraging the flexible and scalable nature of LLMs over other methods. These trade-offs echo the trade-offs provided in the Introduction (Chapter 1) regarding why TMv1 was instantiated with LLMs over other methods. It is being mentioned once more here to answer the natural question arising from the introduction of a programmatic approach utilised to form the GT - "why not just use the programmatic rule-based method over an LLM?"

Dimension	Human	Programmatic	LLMs
Scalability	Manual, limited to domain experts	Template-dependent	Highly scalable generation
Coverage	Expert-bounded	Rule-constrained	Flexible, generalisable
Language Quality	Domain-appropriate	Often rigid, can be unnatural	Natural, fluent
Accuracy Control	Inherent	Deterministic	Requires verification

Table 5.16: Comparison of Methods of Instantiation of CQs.

Thus, while the rule-based method provides a valuable design alternative, LLMs may offer a more practical and sustainable approach from a system-design perspective for dynamic, large-scale KGs.

Figure 5.10 presents an overview of how the LLM and GT filled-in questions and answers were constructed for evaluation.

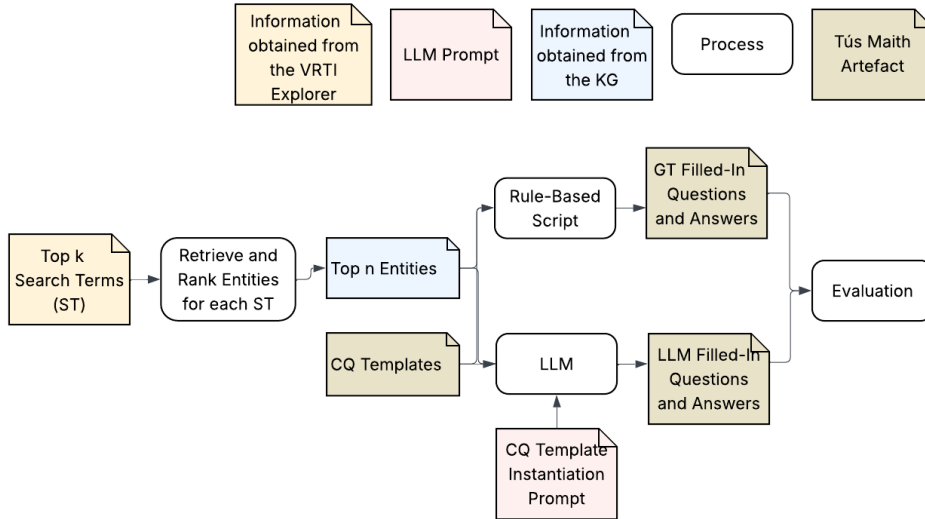


Figure 5.10: LLM and GT Construction for LLM Experiment 2.

5.8.3.3 LLM Inference Setup

Four LLMs were evaluated: two closed-source (GPT-5 [53], Gemini 2.0 Flash [54]) and two open-source (DeepSeek-V3.1 [122], Qwen-2.5-72B-Instruct [123]) at their default settings (temperature=1, top_p=1). Default parameters were employed to assess the models' inherent reliability and provide a realistic, generalisable baseline for their performance in practical, non-technical workflows. GPT-5 was chosen as it represented the most recent generation of OpenAI's large-scale, general-purpose closed-source models, providing a high-performance reference point and continuity with GPT-4o used in LLM Experiment 1 (Section 5.7). Gemini 2.0 Flash was selected as it was also used in LLM Experiment 1 and also to complement GPT-5 by representing a smaller, latency-optimised closed-source model that reflects a contrasting deployment profile focused on efficiency and responsive-

ness. DeepSeek-V3.1 was chosen as a large, high-capacity open-source model, enabling a comparison between state-of-the-art proprietary systems and openly available alternatives. Likewise, Qwen-2.5-72B-Instruct was selected as a comparatively smaller open-source model to examine the effect of model scale and capacity within the open-source category. Collectively, these models were selected to span differences in licensing (closed vs. open), scale, and performance-efficiency trade-offs, providing a representative cross-section of well-established and practically deployable LLMs available at the time. Each model was tested under two zero-shot prompting conditions (both prompts can be found in the Appendix Section A.5.2):

- Basic Provenance: Instructed to return a JSON object of question and answer pairs and cite the source entity URI.
- Full Provenance: The basic prompt but also instructed to also provide the specific JSON keys and values used as evidence.

For each of the 17 batches, 5 runs were executed per prompt condition per LLM (680 full output question and answer sets = 17 batches * 5 runs * 4 LLMs * 2 prompt conditions) generating a total of 24,900 question and answer pairs (total generated for the basic provenance prompt: 12,345, total generated for the full provenance prompt: 12,555), using robust API calls with exponential backoff and comprehensive logging. Because the filled-in questions and answers were generated dynamically by each LLM from batched entities and CQ templates across multiple runs, and because LLM inference is inherently non-deterministic, slight variation in the number of generated filled-in questions and answers was observed between prompt conditions and runs. As mentioned, all of the datasets can be found in the GitHub repository⁴ in the LLM Experiment 2 folder.

5.8.3.4 Evaluation Metrics

To evaluate LLM performance in template filling a set of quantitative metrics spanning four dimensions: (1) accuracy and fidelity of slot filling, (2) completeness of question coverage, (3) robustness against hallucination and prompt non-compliance, and (4) efficiency in runtime. These dimensions collectively answer RQ1_Exp2, which evaluates the overall effectiveness of LLMs in batch template instantiation. RQ1.1_Exp2 is addressed through comparative runs using basic and full provenance requirements, evaluated using semantic similarity metrics, slot-based accuracy metrics, hallucination rate, and runtime. RQ1.2_Exp2 is addressed through automated JSON-based validation of LLM-generated outputs against the GT-dependent semantic similarity classification. Table 5.17 summarises each metric, its definition, and the success thresholds used in this experiment. Full calibration details for semantic similarity thresholds (SBERT [42], BERTScore F1 [43]) and lexical rescue rules are provided in Section 5.8.3.5. Success criteria were established to reflect the dual requirements of output quality and practical utility in question-answer pair curator systems. For quality, thresholds for semantic similarity metrics (SBERT, BERTScore) were derived

⁴https://github.com/mcnamacl/T-s_Maith_Thesis_Repository

from our calibration study (Section 5.8.3.5). For practical utility, targets for coverage (the number of filled-in questions and answers the LLM generates / the total number of possible filled-in questions and answers that could be generated) and LLM hallucination rates were set to ensure the system would meaningfully reduce curator workload: high coverage minimises manual filling effort, while low hallucination rates minimise verification and correction time. Specific thresholds (e.g., $\geq 70\%$ strict completeness, $<5\%$ hallucinations) represent what was judged to be the minimum viable performance for useful deployment.

Table 5.17: Evaluation Metrics and Success Criteria

Category	Metric	Definition / Computation	Success Criteria
Accuracy & Fidelity	Slot-level exact match (Slot Accuracy)	For each required slot in each Q+A ($\{\text{Name}\}$, answer), check JSON value equality after normalisation (lower-case, Unicode, ISO dates). Slot Accuracy = % of slots exactly matched.	$\geq 70\%$ slots match exactly.
	Entity-level all-fields correctness (Full Slot Match)	Binary per generated Q+A, 1 if all slots match exactly, 0 otherwise. Full Slot Match Rate = % of questions fully correct.	$\geq 50\%$ of questions fully correct.
	SBERT similarity (Q+A; answers-only)	Compute embeddings for GT vs. LLM outputs; align sentences using Hungarian 1:1 matching [154] using <code>scipy.optimize.linear_sum_assignment</code> [155]. SBERT Q+A = Mean Cosine Similarity for Q+A, SBERT A = Mean Cosine Similarity for A.	SBERT Q+A ≥ 0.95 , SBERT A ≥ 0.85 .
	BERTScore (Q+A; answers-only)	Token-level semantic match capturing paraphrase/synonym variation. BERTScore F1 Q+A = mean F1 for Q+A, BERTScore F1 A = mean F1 for A.	BERTScore F1 Q+A ≥ 0.90 , BERTScore F1 A ≥ 0.80 .
Coverage	Strict completeness (Strict C.)	Number of fillable questions (slots present in JSON) answered and judged exact/acceptable by SBERT QA thresholds / number of GT Q+As.	$\geq 70\%$ acceptable.

Continued on next page

Table 5.17: Evaluation Metrics and Success Criteria (continued)

Category	Metric	Definition / Computation	Success Criteria
	Lenient completeness (Lenient C.)	As strict, but includes answers rescued via substring/Jaccard similarity ≥ 0.6 or domain-specific exceptions (e.g., DIB URL).	$\geq 85\%$ acceptable.
Robustness	Hallucination rate (Hall. Rate)	Output flagged as hallucination if: low SBERT similarity; no lexical match; not covered by domain-specific exceptions. Hallucination Rate = hallucinations / total generated Q+As.	$< 5\%$ hallucinations.
	Prompt compliance	Automatic checks: (i) valid JSON object, (ii) no code fences/backticks.	No violations.
	Consistency	Across repeated runs with the same prompt, compute distribution of Accuracy & Fidelity and Coverage metrics, and Hallucination Rate.	Narrow confidence intervals and low standard deviation.
	JSON validation agreement	Compare LLM outputs to source JSON (entity URI + values). Report agreement rate, false positives, false negatives against GT.	High agreement with low FP/FN.
Efficiency	Runtime	Mean time taken per run (mm:ss), averaged across batches.	Lower is better, balanced against accuracy.

5.8.3.5 SentenceBERT and BERTScore Calibration

Semantic similarity thresholds were calibrated using a balanced multi-LLM approach to ensure robust evaluation metrics. The calibration set was constructed from a representative pseudorandom subset of 10 entities drawn from the stratified entity pool (5 persons: 2 high, 2 medium, 1 low; 5 places: 4 medium, 1 low), which was passed to each candidate LLM using the basic provenance prompt to generate a pooled set of question-answer pairs for calibration. To ensure robustness and minimise model-specific bias, outputs from all four LLMs (GPT-5, Gemini 2.0 Flash, DeepSeek-V3.1, and Qwen-2.5-72B) were jointly considered. A one-to-one alignment between GT and generated sets was enforced by manually removing questions present in the GT set but absent from the generated outputs. This procedure yielded an initial calibration set of 139 question-answer pairs, which were manually labelled by the author of this thesis as exact ($n=77$), acceptable ($n=51$), or

mismatch (n=10).

As this distribution under-represented mismatches, the calibration set was augmented using a second pseudorandom draw of 10 entities, with one overlapping entity, yielding 19 unique entities in total (removed from the full pool of 192, leaving 173 for the main experiment). From this augmentation, 14 additional mismatches were constructed, comprising six realistic LLM errors (e.g., incorrect county or date), five plausible perturbations (e.g., perturbed years), and three synthetic negatives generated by shuffling answers between questions. To ensure balanced representation across LLMs while incorporating these mismatches, 23 question-answer pairs were sampled per model, where the minimum available count per LLM defined the per-model sample size. Models with larger pools were randomly downsampled to this threshold (resulting in 92 manually-labelled question-answer pairs), and the 14 synthetic mismatches were appended, resulting in a final calibration set of 106 manually labelled examples. These augmentations were used exclusively for calibration and not for subsequent evaluation.

all-MiniLM-L6-v2 [119] was employed to compute SBERT cosine similarity, and roberta-large [121, 55] was used for BERTScore F1, enabling the capture of both coarse-grained semantic similarity and fine-grained token-level variation. Optimal thresholds were identified via grid search with stratified 5-fold cross-validation (reduced from 10-fold due to the limited calibration set size of 106 stratified examples) following [156]. Performance was evaluated using macro-F1 for three-way classification (exact/acceptable/mismatch) [157], implemented via scikit-learn [158]. The classification pipeline incorporated lexical rescue rules prior to the application of calibrated similarity thresholds. The lexical rules are as follows: if the strings are identical, it is considered an exact match; if either string is a substring of the other, it is considered an acceptable match; if the Jaccard similarity between the two strings ≥ 0.6 , is it considered an acceptable match.

The resulting thresholds (Table 5.18) demonstrated appropriate conservatism for each metric type, in which promotion to the exact class requires near-identical semantic similarity after canonicalisation and lexical/domain-specific rescue have been applied.

Table 5.18: Thresholds and Cross-Validation Macro F1 Scores - Pooled Multi-LLM Calibration

Metric	Thresholds	cv-macro-F1
SBERT (Q+A)	$t_{\text{low}} = 0.961, t_{\text{high}} = 0.999$	0.909
SBERT (Answer-only)	$t_{\text{low}} = 0.826, t_{\text{high}} = 0.930$	0.922
BERTScore F1 (Q+A)	$t_{\text{low}} = 0.950, t_{\text{high}} = 0.989$	0.867
BERTScore F1 (Answer-only)	$t_{\text{low}} = 0.938, t_{\text{high}} = 0.968$	0.897

Following [159], a sensitivity analysis confirmed the robustness of the selected thresholds, particularly for SBERT Answer-only, for which 81.9% of nearby threshold combinations remained within ± 0.02 macro-F1 of the optimum. In practical terms, this indicates that small perturbations to the chosen thresholds do not materially change classification performance, and that the SBERT Answer-only operating point lies within a stable region of the macro-F1 landscape rather than at a sharp or fragile optimum.

5.8.3.6 JSON Validation Proxy

A practical consideration was brought to light during the course of the evaluation: namely, that in operational settings, a GT will likely be unavailable with which to compare LLM-generated outputs against. More specifically, the presence of a GT to compare against indicates that there was either a manual or programmatic rule-based instantiation of CQ templates with the entities thereby negating the need for LLMs at all. This consideration directly motivated the design, implementation, and evaluation of a lightweight JSON validation proxy that aimed to classify the generated output as "exact", "acceptable", or "mismatch" (as per the classification using GT used above) via a rule-based classifier. This proxy works in the following manner: (1) the generated source URI exists in the source JSON; (2) the associated entity name matches; and (3) the answer is an exact or acceptable match (per the lexical rules above). If all three checks pass, the generated output is declared as valid and the label assigned to it is the classification provided by check (3). If any fail, the generated output is flagged as needing attention. Check 2 and check 3 occur after string normalisation using the rules that can be found in Appendix A.7.

This lightweight, GT-agnostic JSON validation proxy was inspired by schema/shape validation for structured data (JSON Schema [152]; SHACL [151]) while also resembling programmatic weak/distant-supervision heuristics used when gold labels are unavailable (Snorkel [150], distant supervision [160]); it combined simple lexical/substring/Jaccard checks (cf. string-distance literature [161]), dynamic URI-to-label normalisation, and surface-form matching (cf. entity-linking surveys [162]); following recent advice to "validate the validators" [145], its agreement was quantified with a GT-dependent semantic validator using confusion matrices and standard classification metrics [157]. Specifically, the GT-agnostic generic, basic JSON validation check was compared against the semantic classifier that made use of semantic understanding using SBERT and tailored rescue rules. A key challenge arises when template questions require reasoning over JSON values rather than direct lookup. For example, the question "What century was a person active in?" requires reasoning over the floruit year lower and floruit year upper keys; a direct value match is not possible. Similarly, the question "What type of place is this?" produces valid outputs such as "present day townland," but the source JSON encodes this as a URI (CRM_Type: <https://ont.virtualtreasury.ie/ontology#PresentDayTownland>). Substring and Jaccard similarity checks [161] fail in such cases, despite the semantic match being correct. To handle this JSON encoding issue, dynamic URI handling was introduced: splitting URIs at # or /, normalising the resulting segments, and comparing against the normalised LLM output.

5.8.4 Results

5.8.4.1 Results for RQ1_Exp2 - Accuracy, Compliance, Reliability, Consistency, Efficiency

To evaluate whether LLMs can accurately and reliably fill CQ templates, five runs were conducted using the basic provenance prompt across four LLMs (GPT-5, DeepSeek-V3.1,

Gemini, Qwen-2.5-72B). Each run consisted of 17 batches. Each batch comprised a JSON object with information for ten entities (five people and five places), stratified using the procedure described in Section 5.8.3.1. In total, the runs with the basic provenance prompt generated 12,345 question-answer pairs.

Accuracy, Coverage, and Efficiency Table 5.19 summarises the results and displays the fact that performance varied substantially across models. GPT-5 and DeepSeek both came closest to meeting all success criteria, missing only one criterion each. GPT missing the full slot match rate threshold (0.481, short by 0.019) and DeepSeek missing the lenient completeness threshold (0.841, short by 0.009). GPT achieved the highest SBERT QA (0.975), BERTScore F1 QA (0.986), and completeness (strict = 0.857; lenient = 0.922). DeepSeek-V3.1 performed second best across those criteria and a low hallucination rate (0.003). Its main drawback was longer run time (≈ 52 min). Superficially, looking solely at accuracy and fidelity metrics (slot accuracy, full slot match, SBERT QA, and BERTScore F1 QA), Gemini 2.0 Flash looked as though it performed comparably to DeepSeek and GPT, even outshining them in full slot match rate. However, it is important to take into account that these metrics were only calculated over the total number of its generated question-answer pairs and Gemini’s completeness scores were much lower achieving just 0.559 in strict completeness and 0.614 in lenient. Therefore it was possible to conclude, that while it does not produce many pairs that comply with the completeness criteria, the pairs Gemini does produce are generally accurate. A further critical drawback however, is that in spite of its being the fastest (≈ 4 min per run), Gemini produced an unacceptably high number of question and answer pairs flagged as potential hallucinations at around 12%. Qwen failed the greatest number of criteria at five failures, though three failures were marginal. Its hallucination rate (5.6%) was a little higher than desirable, but runtime was relatively low when compared to DeepSeek and GPT at ≈ 11 min.

Overall, the most accurate models were also the slowest. This created a direct trade-off between quality and speed. Furthermore, Gemini’s low completeness and high hallucination rate meant its speed was not practically useful. Consequently, for a functional system, Qwen may represent a necessary compromise, offering moderate speed and moderate accuracy. However, as time was not an immediate issue for the purpose of evaluating the TMv1 system end-to-end (carried out in User Evaluation 3, Section 5.9), and due to the fact that GPT-5 in general performed the best, it was the LLM of choice utilised in the subsequent TMv1 system prototype end-to-end evaluation.

Hallucination Types In order to make sense of the high rate of hallucinations observed for Gemini, a structured qualitative analysis was conducted over all outputs flagged as potential hallucinations. Rather than treating hallucination as a monolithic phenomenon, the analysis distinguished between different hallucination types as they manifest relative to the strict constraints of the instantiation task.

It was found that a considerable number, 256/283 or 90.5%, of them correspond to answers that were factually grounded but violated the expected output form in a manner that was not detectable by semantic similarity or rescue rules. In these instances, the returned out-

Table 5.19: Performance Metrics by LLM (Mean). Metrics that fall below the success criteria as defined in Table 5.17 in **bold**. N=5, average of 17 batches per run across 5 runs

LLM	Strict C.	Lenient C.	Slot Accuracy	Full Slot Match	SBERT QA	BERT F1 QA	Hall. Rate	Time Taken
DeepSeek	0.797	0.841	0.745	0.502	0.961	0.975	0.003	51:53
Gemini	0.559	0.614	0.766	0.556	0.944	0.980	0.115	3:55
GPT	0.857	0.922	0.713	0.481	0.975	0.986	0.008	40:34
Qwen	0.736	0.814	0.682	0.410	0.941	0.969	0.056	10:52

put contained an answer in URI form like the kind found in run 5, batch 6: "What type of place is Donegal? <https://ont.virtualtreasury.ie/ontology#presentdaycounty>". While technically correct, such outputs are not directly usable within the intended question-answer presentation context and were therefore classified as hallucinations with respect to task compliance. Some however, 16/283 or 5.6%, were malformed e.g. in run 5, batch 15, "What county is a parish/townland/barony located in? south naas". A further hallucination type consists of incorrectly filled template questions with no answer provided such as, "What county is a parish/townland/barony located in?", found in run 4, batch 1. These outputs simultaneously constitute hallucinations and serve as prompt compliance errors (the LLM should only return a question and answer if the answer is not an empty string). These distinctions demonstrate that the observed hallucination rate is not driven primarily by LLM fabrication, but by violations of output expectations defined by the task and the surrounding system. Within the context of TMv1, where generated question-answer pairs must be directly usable by downstream components and end users, such violations are appropriately captured under a single hallucination metric. The high number of non-natural language answers returned suggests that this class of hallucinations may be remedied by explicitly stating in the prompt (beyond "The answer must be a concise fact, not a full sentence. Use the most specific, human-readable field available") to return a natural language answer.

Prompt Compliance and LLM Uptime Reliability Another factor for consideration was prompt compliance. Gemini similarly performed poorly here by adding backtick characters and included the output format name (````json`) resulting in 85 total prompt non-compliance errors across runs. Gemini also had 2 failed batches: Run 1, Batch 2; Run 4, Batch 6.

All runs otherwise completed on first attempt. DeepSeek occasionally triggered API rate-limiting, requiring exponential backoff retries; this behaviour is expected for open-source deployments but important to consider when embedding DeepSeek in a curation system like TMv1.

Consistency To assess consistency, the mean, standard deviation, and 95% confidence interval of the performance metrics across five runs were examined, with the results for mean and standard deviation shown in Tables 5.20 and 5.21. Gemini exhibited low stan-

dard deviations, but this stability was with consistently poor performance, with high hallucination rate. Qwen achieved moderate averages but showed the largest variability (e.g., full slot match mean = 0.410 ± 0.032), indicating weaker reliability. DeepSeek combined strong accuracy with very low variability (e.g., hallucination rate = 0.003 ± 0.002), making it highly reliable across runs. GPT generally achieved the highest mean scores overall (e.g., strict completeness = 0.857 ± 0.015 , SBERT QA = 0.986 ± 0.001) with controlled variability, demonstrating both accuracy and consistency. Together, these results suggest that GPT and DeepSeek are the most reliable scaffolds for CQ template instantiation, while Gemini is unsuitable. Qwen, although less consistent, achieved moderate overall accuracy and could still be usable in contexts where stability is less critical or outputs are subject to human curation. The reliability demonstrated here further motivated the choice of GPT-5 for the domain expert evaluation of TMv1 (Section 5.9).

Table 5.20: Consistency Metrics by LLM (Mean $\pm$ SD) across 5 Runs with Basic Prompt. N=5, average of 17 batches per run

LLM	Strict Comp. Mean $\pm$ SD	Lenient Comp. Mean $\pm$ SD	Slot Acc. Mean $\pm$ SD	Full Match Mean $\pm$ SD	SBERT QA Mean $\pm$ SD
DeepSeek	0.797 ± 0.012	0.841 ± 0.005	0.745 ± 0.012	0.502 ± 0.022	0.961 ± 0.005
Gemini	0.559 ± 0.013	0.614 ± 0.012	0.766 ± 0.011	0.556 ± 0.019	0.944 ± 0.004
GPT	0.857 ± 0.015	0.922 ± 0.004	0.713 ± 0.016	0.481 ± 0.026	0.975 ± 0.002
Qwen	0.736 ± 0.025	0.814 ± 0.032	0.682 ± 0.020	0.410 ± 0.032	0.941 ± 0.014

Table 5.21: BERTScore F1 QA and Hallucination Rate by LLM (Mean $\pm$ SD) across 5 Runs with Basic Prompt. N=5, average of 17 batches per run

LLM	BERTScore F1 QA Mean $\pm$ SD	Hall. Rate Mean $\pm$ SD
DeepSeek	0.975 ± 0.001	0.003 ± 0.002
Gemini	0.980 ± 0.001	0.115 ± 0.016
GPT	0.986 ± 0.001	0.008 ± 0.005
Qwen	0.969 ± 0.002	0.056 ± 0.017

The 95% confidence intervals for means are further presented in Table 5.22 which once again demonstrated that Qwen exhibits the highest variability across metrics. The wider confidence intervals for it (lenient completeness [0.77 - 0.85]) indicated greater uncertainty in its true performance, meaning its results could vary substantially in practice.

5.8.4.2 Results for RQ1.1_Exp2 - Provenance

RQ1.1_Exp2 sought to assess if requesting an LLM to give a full traceback of the JSON keys and values used to fill a template question and answer along with the source URI has an impact on any of the metrics analysed as part of RQ1_Exp2. In order to evaluate this, a further 5 runs across the 4 LLMs were conducted using the full provenance prompt. Paired Wilcoxon Signed-Rank tests [126] were performed on the paired differences

Table 5.22: Consistency Metrics by LLM (Mean and 95% CI). N=5 runs

LLM	Strict Comp.	Lenient Comp.	Slot Acc.	Full Match	SBERT QA	BERT Score QA	Hall. Rate
DeepSeek	0.80 (0.78– 0.81)	0.84 (0.83– 0.85)	0.75 (0.73– 0.76)	0.50 (0.48– 0.53)	0.96 (0.95– 0.97)	0.98 (0.97– 0.98)	0.003 (0.001– 0.006)
Gemini	0.56 (0.54– 0.58)	0.61 (0.60– 0.63)	0.77 (0.75– 0.78)	0.56 (0.53– 0.58)	0.94 (0.94– 0.95)	0.98 (0.98– 0.98)	0.12 (0.10– 0.13)
GPT	0.86 (0.84– 0.88)	0.92 (0.92– 0.93)	0.71 (0.69– 0.73)	0.48 (0.45– 0.51)	0.98 (0.97– 0.98)	0.99 (0.98– 0.99)	0.008 (0.001– 0.014)
Qwen	0.74 (0.70– 0.77)	0.81 (0.77– 0.85)	0.68 (0.66– 0.71)	0.41 (0.37– 0.45)	0.94 (0.92– 0.96)	0.97 (0.97– 0.97)	0.06 (0.04– 0.08)

between the mean of metrics captured as part of Exp2_RQ1 at the LLM x batch level giving 17 points of comparison per LLM and 68 points total for pooled tests (4 LLMs x 17 batches). This pooled test answered the system-level question: does requiring full provenance impact metric outcomes across the set of LLMs evaluated? As nine different metrics were being evaluated, Benjamini-Hochberg False Discovery Rate (FDR) correction [128] was conducted to control the rate of false positives. This pooled analysis was intended to assess system-level effects of provenance requirements rather than to support model-specific inferential claims. As observations are nested within LLMs, results are interpreted as indicative rather than strictly independent samples. The results of the overall tests are presented in Table 5.23. Before correction, statistically significant differences were observed for three metrics: SBERT Answer Similarity, BERTScore Answer F1, and Full Slot Match Rate. After FDR correction, only SBERT Answer Similarity and BERTScore Answer F1 remained significant, with small effect sizes (≤ 0.018). This suggested that requiring full provenance has very little impact on overall accuracy, robustness, and hallucination rate. There was, however, a marked increase in the average time taken for each LLM to complete a run as can be seen in Table 5.24. The greatest increase seen in Gemini at 111.49%. The time taken by DeepSeek and Qwen also increased by over 50% at 62.17% and 58.9% increases respectively and while GPT had the lowest increase, it still was over 15%. Due to the drastic increase in average time taken, when coupled with little to no difference between requiring full vs. basic provenance, and the fact that the JSON proxy requires just the URI, the practical choice, at least for a first pass, is to use basic provenance within TMv1.

5.8.4.3 Results of RQ1.2_Exp2 - JSON Validation Proxy

Table 5.25 shows the comparison between questions that the JSON validator marked as correct (passing all three checks) vs. the questions the GT-dependent Validator identified as correct (SBERT class being exact, acceptable, or rescued) for the basic prompt runs.

Table 5.23: Provenance Analysis Results: Across all LLMs giving 68 points of comparison. BH-FDR correction applied for multiple comparisons (ns = not significant).

Metric	Full	Basic	Difference	Raw P-Value	FDR-Corrected
<i>Semantic Similarity Metrics</i>					
SBERT QA Similarity	0.958	0.955	+0.003	0.2383	0.2681 ns
SBERT Answer Similarity	0.827	0.845	-0.018	0.0028	0.0202 *
BERTScore QA F1	0.976	0.977	-0.001	0.1236	0.2225 ns
BERTScore Answer F1	0.946	0.950	-0.004	0.0045	0.0202 *
<i>Slot-Based Accuracy Metrics</i>					
Strict Completeness	0.751	0.738	+0.013	0.2217	0.2681 ns
Lenient Completeness	0.804	0.798	+0.006	0.7119	0.7119 ns
Slot Accuracy	0.737	0.727	+0.010	0.0905	0.2037 ns
Full Slot Match Rate	0.508	0.488	+0.020	0.0498	0.1495 ns
<i>Robustness Metrics</i>					
Hallucination Rate	0.052	0.045	+0.007	0.1876	0.2681 ns

Table 5.24: Average Time Taken for Basic vs. Full Provenance Runs

Model	Average Basic Prov	Average Full Prov	% Increase
DeepSeek-V3.1	51 minutes 53 seconds	84 minutes 9 seconds	+62.17%
Gemini 2.0 Flash	3 minutes 55 seconds	8 minutes 17 seconds	+111.49%
GPT-5	40 minutes 34 seconds	47 minutes 15 seconds	+16.47%
Qwen-2.5-72B-Instruct	10 minutes 52 seconds	17 minutes 16 seconds	+58.9%

As is demonstrated in the results, the automated JSON validation served as an effective proxy for programmatic GT for direct factual questions but required supplemental approaches for reasoning-based questions. The proxy demonstrated high precision (88-99%) and acceptable recall across all LLMs, making it comparable to the GT validator and therefore likely, operationally viable with appropriate caveats. Namely, additional care taken with questions requiring reasoning. In practice the proxy is conservative: it rarely produces false positives (hence high precision) but can miss reasoning-derived answers (lower recall). Therefore, if implemented within TMv1, the JSON proxy should be used as a first-pass filter in the system and should be implemented so as to flag outputs that require additional, reasoning-aware checks, or human review. It was interesting to note that Gemini’s metrics changed little when reasoning-derived templates are excluded. This can be explained by Gemini’s low completeness, it simply did not produce many question and answer pairs for the questions that required reasoning.

5.8.5 Discussion

This section details how the results translated into actionable design choices for use in the context of implementation of TMv1 and also identifies HCI design implications for TMv1. Ultimately, the results support the core design principle surfaced as part of this thesis, that LLMs are scaffolds, not substitutes.

Table 5.25: JSON Validation Proxy Performance Comparison with GT-dependent Validator (Best values in **bold**)

LLM	Runs	Accuracy	Precision	Recall	F1-Score	Total Valid
All Questions						
DeepSeek	5	0.6344	0.9915	0.6376	0.7760	3137
Gemini	5	0.8352	0.8791	0.9420	0.9094	2528
GPT	5	0.7092	0.9875	0.7146	0.8290	3461
Qwen	5	0.6873	0.9366	0.7154	0.8109	3219
Without Reasoning Questions						
DeepSeek	5	0.7331	0.9910	0.7379	0.8458	2712
Gemini	5	0.8449	0.8832	0.9497	0.9151	2468
GPT	5	0.8044	0.9871	0.8123	0.8909	3049
Qwen	5	0.7454	0.9805	0.7558	0.8533	2814

5.8.5.1 Practical Choice Considerations

There are a number of practical design considerations to be taken into account when using LLMs for the purpose of filling in CQ templates in practice. Namely, in the trade-offs between time taken, monetary cost, accuracy, reliability, and robustness. While generally the most accurate model, GPT-5 is costly and time-consuming. Even so, its accuracy and reliability motivated the use of GPT-5 in User Evaluation 3 (section 5.9). DeepSeek would have been a good second best as it was free at the time (it subsequently changed to cost money), but its uptime was unreliable and it was the most time consuming. While the fastest overall, Gemini produced the least complete output, returned the highest number of potential hallucinations, and generated the highest number of prompt non-compliance errors. If cost and time became an issue, Qwen would be the practical choice out of the LLMs tested when taking into account its time efficiency and moderate results. It is important to note, however, that Qwen performed the most variably across runs which may have an as-yet unknown impact on domain expert curator trust in the automated system – curators may not be able to rely on the quality of its output. For that reason, it was discounted as an option for User Evaluation 3.

5.8.5.2 Effect of the Implications on the TMv1 System

If DeepSeek is excluded due to practical constraints, such as unreliable uptime, in spite of its cost, because of its high performance, GPT-5 emerged as the most appropriate default choice for the TMv1 system prototype for use in User Evaluation 3 (Section 5.9). Qwen represented a viable alternative in scenarios where cost is a significant consideration only if used in tandem with JSON validation proxy due to its variable consistency and moderate results. Superficially, Gemini appears to generate accurate questions and answers. However, it produces an unacceptably high number of potential hallucinations and fills in a low number of possible question and answer templates.

This experiment directly informed a revision of the architecture of TMv1, Figure 5.11, the

original architecture can be seen in the Introduction Chapter, Figure 1.2. Specifically, the figure shows what Stage 2 in the TMv1 system looks like including the lightweight JSON validation step.

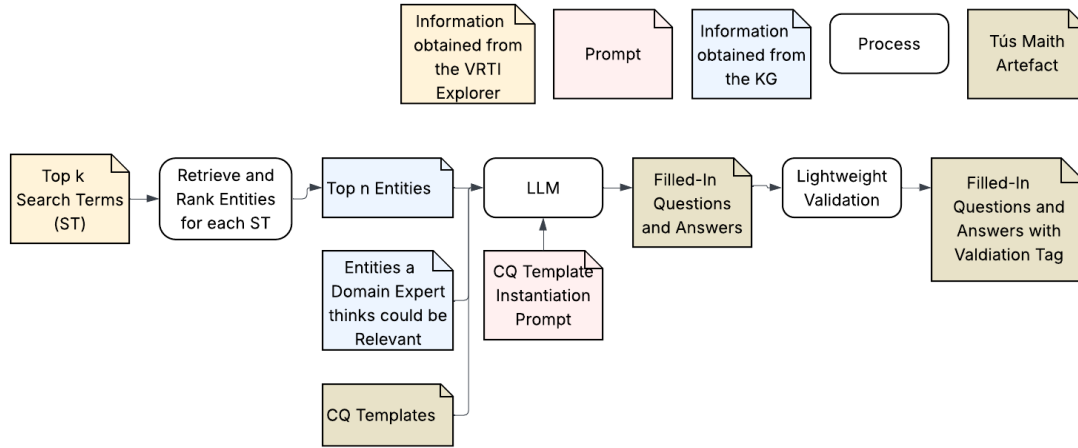


Figure 5.11: Stage two of TMv1 with JSON Validation Step.

Interface-level design implications surfaced by this experiment were applied to the prototype TMv1 interface in Figure 5.12 working with information taken from the VRTI KG and utilised in User Evaluation 3 (Section 5.9). Specifically, the domain expert curator, a VRTI historian, was presented with two key components: (1) the URI referencing the source data used to instantiate the template question and answer, (2) the LLM responsible for the instantiation. Together, these aspects aim to enhance the *explainability* of AI-generated outputs by helping the curators understand not only what was produced, but how it was produced. User-trust of TMv1 as an AI-assisted system was evaluated as part of User Evaluation 3 (Section 5.9).

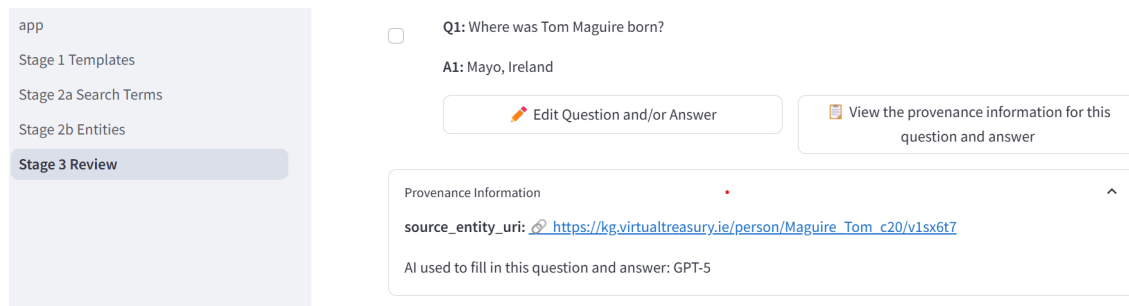


Figure 5.12: TMv1: UI implementation demonstrating the information shown after clicking to view provenance.

5.8.6 Summary of Key Findings

H1_EX2: "LLMs at their default settings (temperature=1, top_p=1) can batch instantiate CQ templates to produce filled-in questions and answers using information about multiple entities" was partially supported. The extent to which can be seen by looking at a more granular level via the experiment's research questions.

RQ1_Exp2: "How effectively can LLMs perform batch instantiation of CQ templates across multiple entities while maintaining accuracy, fidelity, and robustness?" The experiment demonstrated that metrics can vary significantly depending on the LLM chosen. Therefore, careful model selection must occur through consideration of the trade-offs identified (accuracy vs. latency vs. cost). As outlined above, the results of this experiment directly motivated the choice of LLM for User Evaluation 3, the full end-to-end viability evaluation of TMv1 by domain experts – GPT-5.

RQ1.1_Exp2 (Provenance Impact): "How does requiring different levels of provenance evidence (basic: entity URI only vs. full: URI + explicit key/value evidence) affect LLM performance across accuracy metrics (slot-level exact match, entity-level correctness), semantic similarity measures (SBERT, BERTScore), hallucination rates, and efficiency (run times)?" Requesting full provenance (URI + explicit JSON key/value evidence) significantly increased runtime with negligible accuracy gains. This suggests that in practice, TMv1 may operate more functionally requiring just basic provenance (URI). Once again, this outcome directly impacted the design of User Evaluation 3 (Section 5.9) as TMv1 utilised a modification of the basic provenance prompt to examine its viability.

RQ1.2_Exp2 (Validation Proxy): "To what extent can automated JSON validation against source entity data serve as a practical proxy for programmatically created GT when evaluating batch instantiation performance?" A key practical finding was that the JSON validator was effective for direct facts but failed for reasoning-based questions. The impact of this is that it cannot be used as the sole measure of correctness and therefore its role must be as a first-pass filter, with all reasoning-based outputs automatically flagged for curator review. While the validation proxy was not provided as part of the live TMv1 prototype exposed to domain experts, it was particularly useful post-evaluation to assess the structural validity of the LLM-generated filled-in questions and answers generated using CQ templates and entities chosen by domain experts as part of User Evaluation 3 (discussed in more detail in the next section, more specifically, Section 5.9.4.1).

Taken together, these findings support the claim that LLMs are scaffolds for curator-facing KG systems but not substitutes for human curators.

5.9 User Evaluation 3: Assessing the Viability of TMv1

Having established the extent to which LLMs can operate at individual stages of the TMv1 system (Stage 1 - LLM Experiment 1, Stage 2 - LLM Experiment 2) in the previous sections, the final step was to assess the overall viability of TMv1 in supporting domain experts, the intended users of TMv1, create CuQAs. User Evaluation 3 sought to investigate viability of TMv1 supporting user evaluation participants through the triangulation of behavioural, quantitative, and qualitative measures. As such, in this evaluation, viability is treated as an overarching multidimensional concept assessed via a combination of methods rather than one singular construct.

5.9.1 Hypothesis and Research Questions

H1_UE3: TMv1, an instantiation of the Tús Maith framework, is viable for domain experts to create CuQAs.

Viable here refers to the extent to which TMv1 supports end-to-end CuQA creation through an acceptable combination of low perceived workload, manageable time-on-task, sufficient domain expert trust in generated outputs, and the successful production of CuQAs.

As a reminder, the lifecycle of a CuQA is as follows: (Stage 1) CQ template → (Stage 2) filled-in question and answer → (Stage 3) CuQA.

In order to examine the above hypothesis, the following research questions were defined. Similar to LLM Experiment 2, Section 5.8, the research questions have been given different acronyms to differentiate them from the research questions of the thesis. In this case, RQX_UE3, where UE stands for User Evaluation and X refers to the question's number:

RQ1_UE3: To what extent does TMv1 aid domain experts in the efficient creation of publishable CuQAs for the purpose of providing KG exploration starting points for lay users?

The word efficient for the purpose of this evaluation refers to low perceived workload and low time-on-task. This research question can be further split into three secondary lines of inquiry each with distinct methods of assessment as will be presented in Section 5.9.3.1:

RQ1.1_UE3 How usable and efficient do domain experts find TMv1?

RQ1.2_UE3 To what extent does the provision of provenance and model selection transparency provided by TMv1 promote trust in TMv1 within a domain expert?

RQ1.3_UE3 How do domain experts assess TMv1's suitability for supporting the creation of publishable CuQAs?

In this evaluation, "*publishable*" does not imply objective correctness or completeness, but rather denotes domain expert judgement that a filled-in question and answer is sufficient to function as an exploratory entry point for lay users, i.e. meets the criteria to become a CuQA. This was measured behaviourally via selection of filled-in questions and answers and activation of the "publish" button in Stage 3 of the TMv1 prototype system.

Individually, each secondary research question examines a specific contributing aspect of overall viability rather than viability in isolation.

5.9.2 Experimental Setup

5.9.2.1 TMv1 Prototype

Utilising the best LLM identified by LLM Experiment 2 (Section 5.8), GPT-5 [53], and the prompt protocol identified as the most promising, and remaining consistent with the number of CQ templates requested in LLM Experiment 1 (Section 5.7), 20 candidate CQ templates were generated (which can be found in the thesis GitHub repository⁵ in folder User Evaluation 3/T-s_Maith_TMv1/data). Specifically, GPT-5 was given a few-shot prompt with a more compact version of the VRTI Ontology (Appendix A.5.3). Both the prompt and ontology subset were updated to reflect lessons learned from LLM Experiment

⁵https://github.com/mcnamacl/T-s_Maith_Thesis_Repository

1 and LLM Experiment 2 (the prompts should be extremely explicit) and the changing nature of the VRTI KG (namely, some classes and properties were removed, others added). For clarity and accessibility during the evaluation, LLM involvement was described to participants using the term "AI"; however, for analytical consistency, all references to AI-generated content in this evaluation refer specifically to outputs produced by an LLM within TMv1.

Of the 20 candidate CQ templates generated, six were selected by the author to be presented to user evaluation participants during the evaluation. A reduced set was intentionally used to streamline the evaluation procedure as the aim of the study was to assess the overall viability of TMv1 rather than to exhaustively cover all of the possible CQ templates. Such a reduction was considered justifiable because LLM Experiment 1 had already demonstrated the LLM’s capability to generate viable CQ templates. The selected six CQ templates were chosen to form a stratified sample, comprising three person-related (CQ1 - CQ3) and three place-related questions (CQ4 - CQ6).

Table 5.26: CQs Presented to Domain Experts

CQ Label	Question Text
CQ1	Where was a person born?
CQ2	Which sources list a person?
CQ3	What century is a person associated with?
CQ4	What larger area does a place fall within?
CQ5	What era is a place associated with?
CQ6	Which sources refer to a place?

5.9.2.2 Participants

In order to ensure that the evaluation complied with Trinity College Dublin research ethics guidelines, an ethics application was designed and submitted, then subsequently approved by the School of Computer Science Research Ethics Committee. The full ethics application can be found in the Appendix [A.4.4](#).

Sampling Strategy This evaluation required domain expertise in history and familiarity of the VRTI KG in order to be able to evaluate the viability of TMv1 in supporting historians in the generation of CuQAs. Therefore, the sampling pool was limited to historians working with or adjacent to the VRTI project. Following Guest et al. [163], who demonstrated saturation in homogeneous qualitative samples by $n = 6-12$, this evaluation prioritised rich expert input to indicate viability, given the limited participant pool. Six participants responded in total from the recruitment email (Appendix [A.1.4](#)) sent to historians working on the VRTI project and historians working on a project closely adjacent to the VRTI project. One of the six participants was DE2 from User Evaluation 2, Section 5.6, and another was DE3 from LLM Experiment 1, Section 5.7. While these two participants could be considered more familiar with the concept and intended purpose of CuQAs than the other participants, neither had interacted directly with TMv1 prototype

system prior to this evaluation. Specifically, one participant’s prior involvement was limited to evaluating CQ templates (LLM Experiment 1, Section 5.7) (DE3), while the other had contributed to the development of CuQAs used in a separate user evaluation (User Evaluation 2, Section 5.6) (DE2). As neither participant had prior exposure to TMv1 end-to-end or the interface, their earlier involvement was not expected to influence their assessment of TMv1’s usability, workload, or perceived viability; rather, their familiarity with the underlying concepts can be considered beneficial in enabling informed, critical evaluation of TMv1’s suitability for CuQA creation. The results of the evaluation were anonymised; therefore, it was not possible to identify which of the six participants DE2 and DE3 corresponded to. As such, for the purpose of interpreting the results of the evaluation, the participants were henceforth referred to as PX, where X corresponds to the participant’s number.

Inclusion Criteria As detailed in the above section, due to the specialised nature of the evaluation, strict inclusion criteria were employed. The criteria were as follows: Historical domain experts familiar with the VRTI project and KG. Anyone who did not fit the inclusion criteria was excluded.

5.9.3 Evaluation Procedure

Each evaluation session followed a structured procedure consisting of an introductory briefing, task-based interaction with the TMv1 prototype, and a post-task questionnaire. Sessions were conducted online via Microsoft Teams individually and the task-based interaction with TMv1 lasted approximately 11 minutes on average.

Introduction and Briefing At the beginning of each session, participants were briefed on the purpose and scope of the evaluation. The aim of the evaluation was explained as assessing the viability of TMv1 for supporting the creation of CuQAs for the purpose of providing exploratory entry points for users wanting to begin interaction with the VRTI KG. Participants were informed that the focus of the evaluation was on the generation of simple CuQAs that can act as exploratory starting points into the KG (i.e. ones that map to SPARQL queries and whose attributes lie in the immediate semantic neighbourhood of an entity) as opposed to other question types, such as interface-dependent CuQAs (e.g., CuQAs that lead to interface features) and complex CuQAs (multi-hop and/or compound CuQAs).

Participants were encouraged to provide feedback as they made their way through the stages of TMv1, on the quality of generated outputs, and their level of trust in the system throughout the session.

Regarding how the use of AI was conveyed to the participants, the script created for the evaluation specifically told the participants that: "This interface is an early-stage workflow that makes use of AI and is designed to support domain experts in generating simple featured questions and answers that can act as entry points into the VRTI knowledge graph" (once again, the full script can be found in the Appendix A.3.1).

During stage one, the choosing of CQ templates, as detailed below, the UI tells the participants that the CQ templates they are presented with have been generated by AI. Furthermore, proceeding from Stage 2 to Stage 3 (the filling in of chosen CQ templates to create filled-in questions and answers with chosen entities to be reviewed by the domain experts), the UI presented to the participants the fact that this process was being carried out by AI, Figure 5.13.

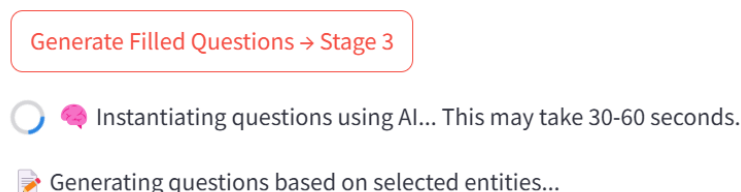


Figure 5.13: Conveyance of the use of AI to domain experts.

Task Execution using TMv1 Participants then completed a task-based evaluation using the TMv1 prototype. During task execution, behavioural interaction data were recorded per participant, including total task completion time, time spent at each stage of TMv1, interactions with provenance and model transparency features, the filled-in questions and answers produced, number of edits made to system-generated filled-in questions and answers, and the number of filled-in questions and answers selected to become CuQAs. Participants progressed through the following stages:

1. **Stage 1: Choosing CQ Templates.** Participants were presented with six CQ templates (three person-related and three place-related) as seen in Figure 5.14. They were asked to review the templates and select four that they considered suitable for instantiation.

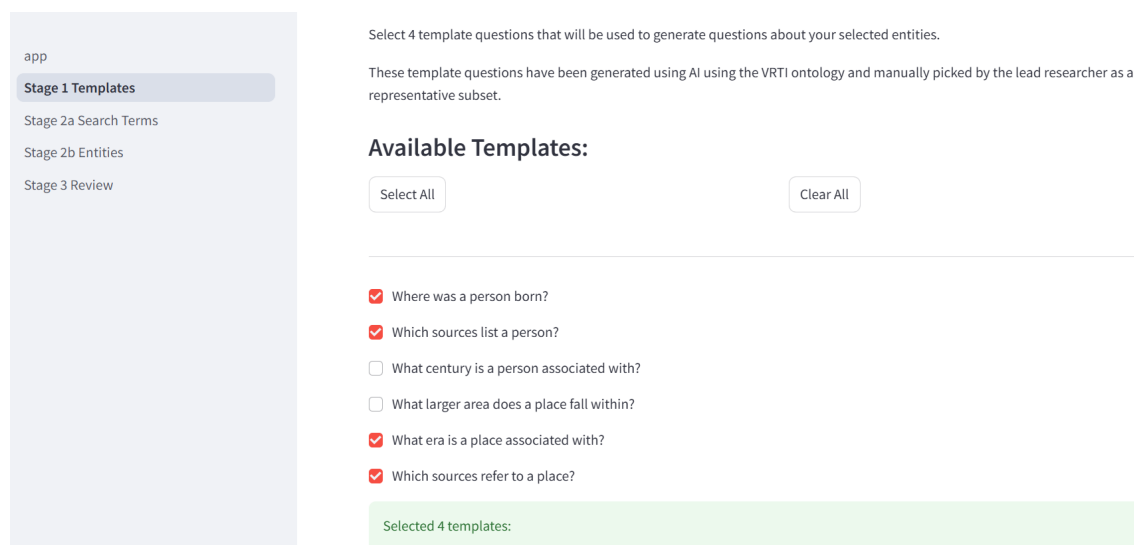


Figure 5.14: Stage 1: Choosing of CQ Templates.

2. **Stage 2: CQ Template Instantiation.** Participants firstly selected one search term from a set of the top five search terms public users had been searching over the VRTI Explorer [37] (Figure 5.15).

Figure 5.15: Stage 2a: Choosing of search term to generate entities.

Once they had chosen a search term, they were presented with the top five person and top five place entities ranked using the entity ranking algorithm (Appendix A.6) created as part of LLM Experiment 2, Section 5.8 (Figure 5.16). From the listed entities, they were asked to choose two person entities and two place entities that would be used to fill in the question templates they had previously chosen.

Figure 5.16: Stage 2b: Choosing of two person and two place related entities to fill in the previously chosen CQ Templates.

Once they had chosen four entities, participants clicked to instantiate the template with their chosen entities as shown in Figure 5.17. As explained previously, this process along with the script provided the participants with the information that the system they were utilising (TMv1) was an AI-assisted one.

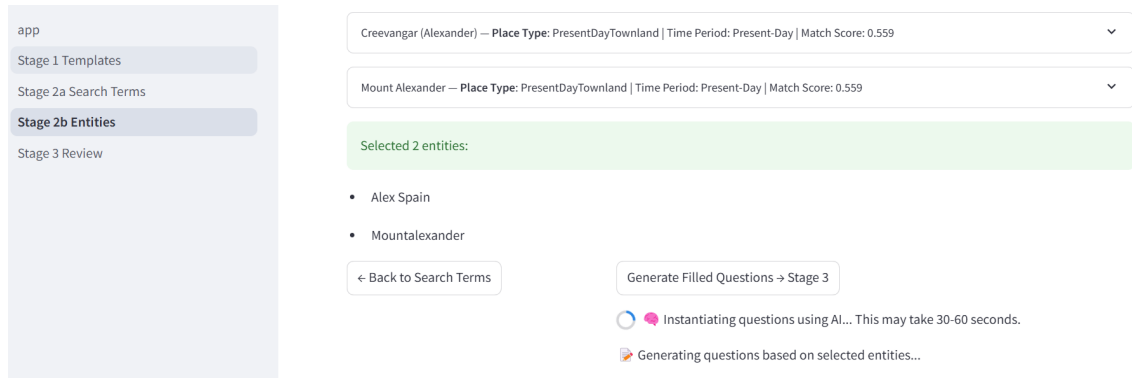


Figure 5.17: Stage 2: CQ Template Instantiation.

3. **Stage 3: Expert Curation.** Participants reviewed the resulting filled-in questions and associated answers generated by the system and were also given the option to edit them. A button also allowed them to view information about how the question-answer was generated (that is the provenance). They then decided which, if any, of the filled-in questions and answers could serve as CuQAs for end users. An example of what was shown to participants when they selected a filled-in question and answer can be seen in Figure 5.18. Interaction with provenance information and model transparency features was observed and logged by the prototype. Participants were encouraged to explain their reasoning when engaging with these features. If the participant chose filled-in questions and answers that they deemed publishable, once they were finished reviewing, they were instructed to click the "Publish" button. This was in order for the system to be able to log the number of publishable filled-in questions and answers (CuQAs), and also to end the timer for stage 3. The "Finished Reviewing" button was there to end the timer in the event that the participant chose no filled-in questions and answers to be publishable.

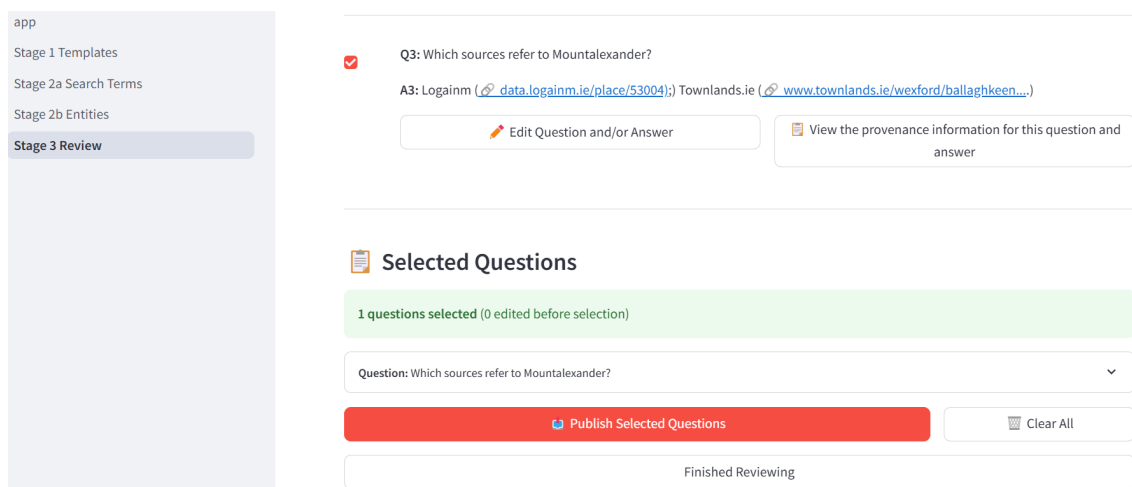


Figure 5.18: Stage 3: Reviewal of instantiated questions and answers.

Post-Task Measures Once the evaluation task was completed, the online call between researcher and participant was ended. The participants then completed a post-task ques-

tionnaire. The questionnaire can be seen in Table 5.27 with each question mapped to its respective research question and questionnaire construct, where construct refers to a measured concept [164] (System Trust, TMv1 Suitability, Workload). Each question was measured on a 1 - 5 Likert scale [130] with 1 being "not at all", and 5 being "very". For the purpose of the evaluation, TMv1 was referred to as a workflow.

Table 5.27: Questionnaire Question Mapping to Research Questions

Research Question	Construct	Q. ID	Question
RQ1.1_ UE3: How usable and efficient do domain experts find TMv1?			
RTLX	Workload	TLX1	How mentally demanding was the workflow? Consider the amount of thinking, reasoning, and concentration required to complete each stage. (Mental Demand)
		TLX2	How rushed or time-pressured did you feel while completing the workflow? (Temporal Demand)
		TLX3	How successful do you feel you were in producing questions and answers using the workflow? (Performance)
		TLX4	How hard did you have to work to complete the workflow to your satisfaction? (Effort)
		TLX5	How frustrated or uncertain did you feel while using the workflow? (e.g., discouraged, annoyed, or unsure about what to do next). (Frustration)
RQ1.2_ UE3: To what extent does the provision of provenance and model selection transparency provided by TMv1 promote trust in TMv1 within a domain expert?			
S-TIAS/Trust	System Trust	TR1	I am confident in the system.
		TR2	I can trust the system.
		TR3	The system is reliable.
		TR4	I felt confident in the system’s output because I could see how the question and answer was generated.
RQ1.3_ UE3: How do domain experts assess TMv1’s suitability for supporting the creation of publishable CuQAs?			
TMv1 Suitability	Complexity	V1	How complex or conceptually difficult did you find the workflow?
	Output Quality	V2	To what extent do you believe this workflow can produce questions and answers that fit within the scope of the VRTI knowledge graph?
		V3	To what extent do you believe this workflow can generate questions and answers that lead users to information-rich areas of the knowledge graph?

Continued on next page

Table 5.27: Questionnaire Question Mapping to Research Questions (continued)

Research Question	Construct	Q. ID	Question
		V4	How likely is this workflow to produce questions and answers suitable for the VRTI's intended public audience?
		V5	To what extent do you believe the workflow can support the generation of clear, well-phrased curated questions and answers?
		V6	How confident are you that this workflow can produce questions and answers answerable by the VRTI knowledge graph?
	Overall Viability	V7	Overall, how viable do you believe this workflow is for generating questions and answers suitable as entry points into the VRTI knowledge graph?
Contextual & Qualitative Items			
Prototype Context	Grounding	G1	To what extent did the limitations of the current prototype affect your evaluation?
Qualitative Feedback	Open-ended	OE1	Please explain your reasoning behind your previous rating.
		OE2	Please provide any comments you have about the workflow and/or your experience in the text box below. (Do not input any identifiable information about yourself)

5.9.3.1 Data Collection Instruments

Overview To evaluate the viability of TMv1, data was collected using a combination of behavioural interaction metrics, a moderated concurrent verbalisation protocol [116], and a post-task questionnaire. These instruments were selected to triangulate efficiency, system trust, and perceived task suitability, corresponding respectively to RQ1.1_UE3, RQ1.2_UE3, and RQ1.3_UE3, which together contribute to the overall assessment of TMv1 viability.

Behavioural Interaction Metrics During task execution, a set of behavioural metrics was captured to provide objective indicators of efficiency, engagement with provenance information (system trust), and candidate CuQAs generated (TMv1 suitability). These metrics were logged automatically by the TMv1 prototype and can be seen in Table 5.28 below.

These measures were used to triangulate self-reported workload, trust perceptions, and TMv1 suitability obtained via the questionnaire. More specifically, the stage-level time-on-task was recorded as the following: Stage 1: start time - entry of participant ID → end time - submission of Stage 1 CQs. Stage 2: start time - once the entity is chosen → end

Table 5.28: Behavioural Metrics for Evaluating TMv1 Interaction

Metric	Description
Task completion time	The total time in seconds taken by the user to complete the entire task using TMv1 from start to finish.
Stage-level time-on-task	The time spent by the participant on each stage of TMv1.
Provenance interaction	The count of user interactions with provenance-related controls i.e. inspecting the provenance of LLM-generated filled-in questions.
Editing behaviour	The number of edits made by the participant to the LLM-generated filled-in questions.
Output yield	The count of filled-in questions and answers that are ultimately selected by the participant to become (CuQAs) and CuQA percentage: count of CuQAs / total number of filled-in questions and answers. This metric serves as the primary behavioural indicator of TMv1’s output-oriented viability, interpreted alongside workload and trust measures to assess overall viability.

time - clicking of "Generate Filled Questions". Stage 3: start time - entry into the reviewal → end time - clicking of "Publish Selected Questions" or "Finished Reviewing".

Think-Aloud Protocol A guided moderated concurrent verbalisation protocol [116] was employed in order to understand from a qualitative perspective the behaviours noted in the above section. The purpose of this protocol was explanatory and triangulatory rather than exploratory. In other words, it was used to understand participants’ reasoning, trust calibration and decision-making during interaction with TMv1, and to contextualise the observed interaction patterns. The script used as part of the protocol can be found in Appendix A.3.1 and each session was audio recorded. Participants were encouraged to articulate their thoughts as they progressed through each stage of TMv1, including their use of provenance information and their editing of candidate CuQAs. The moderator’s role was primarily procedural in guiding participants through the scripted stages, but also minimally probative, seeking clarification where necessary to interpret specific observed behaviours (e.g., why provenance was viewed). The recordings of each session, post-evaluation, were transcribed using a mix of automated [117] and manual (transcription by the author) methods. Given the guided and interactive nature of the evaluation, the transcriptions were not subjected to formal thematic analysis or inductive coding, as the interactive moderation precludes reliable identification of unprompted cognitive strategies.

Questionnaire Instruments The participants were asked a set of 20 questions (see Table 5.27) post-evaluation via Google forms. These questions consisted of a combination of 1 - 5 Likert scale [130] questions and open-ended questions. More specifically, the questionnaire included measures of perceived workload using an adapted version of the NASA Raw Task Load Index (RTLX) [113], trust in the system using a short-form Trust

in Automated Systems scale (S-TIAS) [115], an assessment of the suitability of TMv1 using a custom question set, and an assessment of the quality and suitability of the generated filled-in questions for use as CuQAs. The latter employed a custom question set adapted from the one used in LLM Experiment 1 (Section 5.7). An outcome of LLM Experiment 1 directly informed this adaptation. Specifically, no inter-rater reliability (ICC) was observed for relevance judgments between two domain experts (DE1 and DE3) when assessing CQ templates produced by LLMs. In response, the relevance criterion was reformulated using a clearer rubric that decomposed relevance into three sub-components: topical fit, exploratory value, and audience match. Open-ended items allowed participants to provide additional comments on their experience using TMv1. Following [114], RTLX was modified to remove the question relating to physical demand as it was not deemed relevant to the evaluation. The other five demands (mental, temporal, performance, effort, and frustration) were captured however, by altering the wording of the questions slightly to align them more closely with the evaluation.

5.9.4 Results and Implications

The following section first presents the validation step performed over the generated fill-in questions and answers by all participants. It then discusses the results obtained from the evaluation and the implications of the results on the viability of TMv1 in supporting domain experts in the generation and selection of CuQAs. The full set of results can be found in the thesis GitHub⁶ in the User Evaluation 3 folder.

5.9.4.1 Validation Step

While not carried out during the evaluation to preserve focus solely on the three stages of Tús Maith as instantiated by TMv1, post-evaluation, all filled-in questions and answers generated per participant were subjected to the JSON validation proxy detailed in LLM Experiment 2, Section 5.8.3.6. The result confirmed a 100% validation rate for the filled-in questions that did not require reasoning to produce their answer (non-reasoning filled-in questions and answers = 32 out of the 38 total). As previously demonstrated in LLM Experiment 2, the JSON validation proxy achieved 99% precision for GPT-5 (the LLM used here) when compared against GT semantic similarity validation, indicating that structurally invalid or mismatched filled-in questions and answers were highly unlikely to be present. The six remaining filled-questions (the ones that utilised the C3, **What century is a person associated with?**), were manually inspected by the author of this thesis to ensure correctness based on the entity information. All 6 satisfied the three JSON validation proxy checks, confirming their structural validity despite requiring reasoning to generate. This ensured that participant evaluations were based on technically valid outputs. Judgements concerning historical adequacy and suitability were intentionally delegated to domain expert selection and curation, consistent with the system-level scope of this evaluation.

⁶https://github.com/mcnamacl/T-s_Maith_Thesis_Repository

5.9.4.2 Analytic Framing and Handling of Prototype Constraints

During analysis of the questionnaire results, one participant (P3) exhibited a markedly different response pattern across trust, perceived workload, and perceived suitability measures for TMv1. Inspection of the contextual grounding item (G1) and accompanying qualitative feedback indicates that this participant's evaluation appeared to be strongly shaped by perceived limitations of the current TMv1 prototype instantiation. This focus seemed to be directed toward the prototype rather than the underlying Tús Maith framework as instantiated in TMv1. Specifically, P3 noted that the restricted range of entities and question types made it difficult to assess processes of choice, evaluation, and refinement that are structurally supported by TMv1 but only partially expressed in the prototype used in the evaluation. This suggests a construct misalignment in which the participant primarily evaluated prototype expressiveness and coverage rather than the viability of the TMv1 instantiation as a realisation of the Tús Maith framework, as intended by RQ1_UE3. More specifically, the author of this thesis defines construct misalignment as a mismatch between the intended evaluative target (TMv1 viability) and the participant's interpreted target (prototype expressiveness).

Limitations of prototype fidelity are a known characteristic of early-stage framework instantiations [165]. As this concern was explicitly anticipated and instrumented via the G1 grounding question, in this chapter results are reported both including and excluding P3. This dual reporting supports transparent interpretation of TMv1's perceived viability while acknowledging the influence of implementation constraints on participant responses. While one additional participant also reported a high impact of prototype limitations (G1 = 5), their quantitative ratings and qualitative commentary remained aligned with an assessment of TMv1's intended instantiation rather than a wholesale evaluation of prototype constraints. As such, only P3 exhibited evidence of construct misalignment.

5.9.4.3 Quantitative Overview of Questionnaire Constructs

The overall results for each questionnaire construct can be seen in Figure 5.19. It is important to note when observing the figures and interpreting the results that, for System Trust and TMv1 Suitability in creating publishable CuQAs, the closer to 5 the better. This is in contrast to Workload, where a score closer to 1 is better. As previously mentioned, the questionnaire required responses on a 1 - 5 Likert scale, where 1 indicates "not at all" and 5 indicates "very", therefore 3 is considered neutral (neither better nor worse). Given the small expert sample size ($n=6$), results are reported descriptively without inferential statistical testing.

As is evident from Figure 5.19, the average score for System Trust and TMv1 Suitability are both above 3, at 4 and 3.95 respectively. This indicates that overall, participants felt that TMv1 was trustworthy and that it was a suitable way to generate publishable CuQAs. Average Workload was below 3 at 2.1, suggesting that while TMv1 was not considered very challenging, there was a low to moderate workload felt by participants. It is possible to see moderate to high degree of variability of the average scores as indicated by the standard deviation error bars.

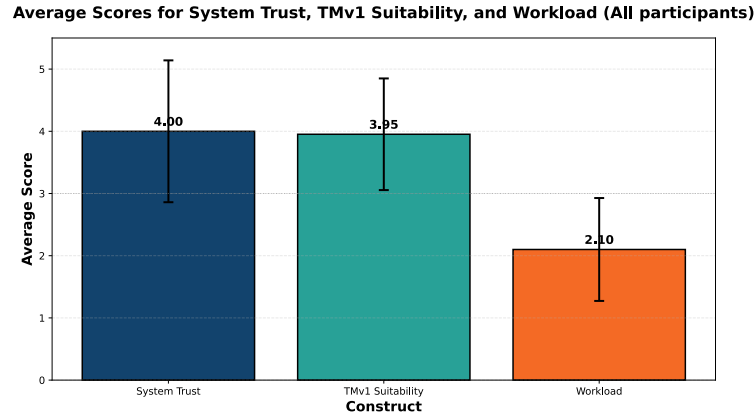


Figure 5.19: Questionnaire construct averages and standard deviations across all 6 participants where for System Trust and TMv1 Suitability, the higher score the better, and for Workload, the reverse is true; the lower the better.

A similar but less variable view of the overall results can be seen in Figure 5.20 which presents the overall averages without P3. System trust and TMv1 Suitability increase from 4 to 4.4, and 3.95 to 4.29 respectively. Workload decreased from 2.1 to 1.8 indicating a low perceived workload. The decreased variability demonstrates that the averages without P3, cluster more tightly together indicating closer agreement between the other five participants.

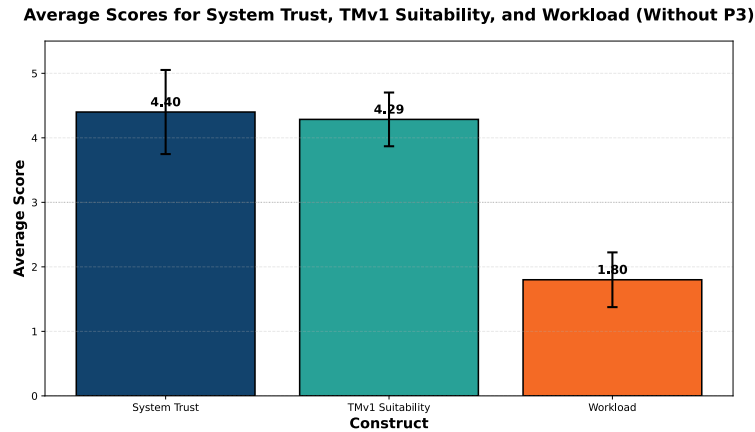


Figure 5.20: Questionnaire construct averages and standard deviations without P3 where for System Trust and TMv1 Suitability, the higher score the better, and for Workload, the reverse is true; the lower the better.

These aggregate patterns provide a descriptive overview of participant responses across the three measured constructs. Interpretation of these results in relation to usability, trust, and suitability is carried out through triangulation with behavioural interaction metrics and qualitative evidence in the following research question-driven subsections.

5.9.4.4 Workload (RQ1.1_UE3)

Workload and efficiency were assessed through triangulation of self-reported workload (modified RTLX), objective time-on-task measures, and qualitative commentary captured

during task execution. Together, these measures provide complementary perspectives on the cognitive and temporal demands imposed during the use by participants of TMv1. Across participants, average workload scores were low to moderate (mean = 2.1 on a 1-5 scale), indicating that TMv1 was not perceived as cognitively demanding overall. This perception is supported by task completion times, with participants completing the end-to-end task of curation in an average of approximately 11 minutes (mean = 680.44 seconds), suggesting that CuQA generation could be executed efficiently within a single session. As shown in Table 5.29 (granular per participant scoring) and Table 5.30 (overall average and standard deviation), variability in workload ratings broadly aligned with differences in interaction patterns. For example, P3 reported the highest workload score (3.6), yet completed the end-to-end task in a comparable time to other participants, indicating that perceived difficulty may have stemmed more from evaluative uncertainty than from interaction complexity. In contrast, P5 exhibited the longest total completion time, largely attributable to extended engagement in Stage 3, which was accompanied by a moderate workload rating rather than frustration or task breakdown.

Table 5.29: Participant Average Workload Scores and Task Completion Times in seconds.

PX	Average Workload Score	Stage 1	Stage 2	Stage 3	Total Time
1	1.2	97.067	129.616	336.181	562.864
2	1.8	116.561	122.699	475.059	714.319
3	3.6	127.453	161.928	259.229	548.610
4	1.6	79.822	132.080	360.021	571.923
5	2.2	121.018	215.481	848.643	1185.142
6	2.2	40.950	166.785	292.053	499.788

Table 5.30: Average Workload and Task Completion Times (Mean $\pm$ Std Dev)

Workload	Stage 1	Stage 2	Stage 3	Total Time
2.1	97.15 $\pm$ 29.79	154.76 $\pm$ 31.77	428.53 $\pm$ 199.65	680.44 $\pm$ 235.05

5.9.4.5 System Trust (RQ1.2_UE3)

Average reported trust of TMv1 triangulated with behavioural and qualitative information can be seen in Table 5.31. Provenance views refers to the number of times the participant clicked "View Provenance" for a generated filled-in question and answer and the reason for viewing is the reason the participant gave when asked why they chose to click to view the provenance information.

The results suggest that, in general, the participants trusted TMv1 with only one of the six (P5) explicitly stating that among the reasons they viewed the provenance information was out of distrust of AI involvement. More specifically, when asked why they clicked to view provenance, P5 stated that, "especially knowing that there is an LLM involved," that they "like the clarity of knowing, being able to trace back the information, [and] being able to check where that came from and who this person is." This was reflected in P5's

Table 5.31: Participant Trust Scores, Provenance Viewing Behaviour, and Perceived Sufficiency

PX	Average Trust	Provenance Views	Reason for Viewing	Prov. Sufficiency*
1	5.0	0	NA	NA
2	5.0	3	Curiosity	5
3	2.0	0	NA	NA
4	4.0	1	Verification	4
5	3.5	4	Distrust/Verification	5
6	4.5	4	Curiosity	4

*Response to: "The provenance information was sufficient for me to verify the accuracy of the filled-in questions and answers." (1-5 scale)

lower quantitative average trust score of 3.5. Their open-ended questionnaire comments also align with their quantitative responses and transcript comments where they positively stated that they "liked the workflow a lot and especially appreciated the transparency in the answers".

The behavioural data revealed that interaction with provenance features was selective rather than uniform. Participants who reported very high trust (e.g., P1 and P2) engaged minimally or not at all with provenance information, often citing curiosity rather than verification when they did so. This suggests that transparency features also support trust calibration rather than blind acceptance.

In spite of P4 reporting a lower average trust of the system (4.0) when compared to the sample median (4.25), their qualitative comments from the questionnaire and transcript suggested interest in the provenance for verification purposes rather than distrust. However, they did note specifically that they were "particularly happy to see an option to check the provenance of the question-answer combinations generated" and that they "believe this is crucial to user trust in the workflow". Taken together, these mixed signals indicate that interaction with provenance features cannot be cleanly classified as solely indicative of either trust or distrust. Instead, provenance use functioned as a trust-calibration mechanism. In some cases, participants engaged with provenance out of interest or curiosity to better understand how answers were generated, while in others it was used to verify or challenge system outputs when uncertainty or scepticism was present. This dual function helps reconcile differences between transcript comments and post-task questionnaire responses, which capture participant reasoning at different moments and under different levels of reflection.

P3's low average trust score did not directly suggest distrust of AI involvement as they did not choose to view provenance. Rather, it appears to reflect broader scepticism driven by perceived prototype limitations rather than by opacity or mistrust of the underlying generation process.

5.9.4.6 TMv1 Suitability (RQ1.3_UE3)

Overall, TMv1 was perceived as moderately to highly suitable to produce publishable CuQAs by most participants as can be seen in Table 5.32 where average suitability refers

to the average of the scores per participant for the Suitability construct. This was further reflected by the behavioural output as all six participants chose at least two filled-in questions and answers that they felt could act as entry points into the VRTI KG (questions published), turning those filled-in questions and answers into CuQAs. The number of edits made to the filled-in questions and answers was also captured by the prototype and is represented in Table 5.32. Within Table 5.32, "Total No. of Filled-In QAs" refers to the number of filled-in questions and answers generated by the prototype based on the participants chosen CQ templates and entities, "No. of CuQAs" indicates the number of filled-in questions and answers selected by the participant for publishing, and "Conversion Rate" presents the percentage of filled-in questions and answers the participants felt were suitable to become CuQAs.

Table 5.32: Participant TMv1 Suitability Scores and Output Metrics.

PX	Average Suitability	Total No. of Filled-In QAs	No. of CuQAs	Conversion Rate (%)	Number of Edits Made
1	4.86	6	6	100	5
2	4.57	8	4	50	0
3	2.29	8	8	100	0
4	4.14	8	2	25	0
5	4.00	8	4	50	0
6	3.86	6	6	100	0

Only one participant (P1) decided to edit their questions and answers before deeming them suitable to become CuQAs. These edits were primarily presentational and focused on the ordering of entity names in the question text (preferring forename-surname over the system-generated surname-forename format).

While the TMv1 prototype logged all filled-in questions and answers generated by the LLM on a per-participant basis and recorded the number of edits and publication actions at the participant level, it did not retain artifact-level records (e.g., filled-in questions and answers chosen as publishable or edited filled-in questions and answers), as the evaluation focused on overall workflow viability rather than on content-level assessment of CuQAs. As such, editorial intervention is evidenced here through transcript-grounded commentary paired with participant-specific system outputs.

For example, P1 noted: "Okay, so I just edit them if I think that they should be worded differently ... it may probably make more sense to have it forename, surname." A candidate question presented to P1 by the system was "Where was Spain, Alex born?", which the participant indicated they would reformulate as "Where was Alex Spain born?" In this case, only the wording of the question was targeted; no changes to the answer content were described or observed.

In spite of expressing the strongest concerns regarding prototype limitations, P3 ultimately selected the highest number of filled-in questions as suitable for publication as CuQAs. In their commentary, P3 emphasised that while some questions constituted stronger starting points than others, none were illogical or misleading, and all were deemed capable

of supporting initial exploration for novice users, such as first-year or secondary school students. This apparent divergence between questionnaire ratings and selection behaviour suggests that P3’s evaluation focused less on the quality or appropriateness of the generated outputs, and more on the limited expressive range of the prototype instantiation for supporting comparative judgment and refinement. Importantly, this indicates that even under a critical evaluation shaped by prototype constraints, TMv1’s outputs were still perceived as meeting the minimal criteria for use as exploratory entry points into the VRTI KG.

Conversion rates ranged from 25% to 100% across participants, indicating variability in how selectively domain experts judged filled-in questions suitable for publication (conversion to CuQAs). This variability suggests differences in individual acceptance thresholds rather than uniform endorsement of all generated outputs.

It must be noted however, that several participants expressed a desire for more complex and semantically rich questions than those typically produced by the current prototype. While the generated questions were consistently described as "simple" and "plain," participants emphasised the importance of supporting more specific and historically expressive queries, such as relational or temporal aggregations (e.g., office-holding over a defined period or life-event trajectories). This feedback suggests that, while TMv1 meets minimal sufficiency requirements for exploratory entry points, expert engagement and sustained use may depend on the framework’s ability to support higher-complexity CuQAs as the underlying KG matures.

5.9.4.7 Triangulated Interpretation and Implications for TMv1 Viability

Taken together, the questionnaire results, behavioural interaction metrics, and qualitative feedback indicate that TMv1, as an instantiation of the Tús Maith framework, was perceived as viable by the domain experts for the creation of publishable CuQAs. Questionnaire measures suggest generally high trust, low perceived workload, and positive assessments of suitability, with divergence in one case attributable to prototype expressiveness rather than conceptual suitability.

Behavioural data reinforce this interpretation, particularly through consistent task completion, engagement with provenance features, and the selection of publishable CuQAs across participants. Importantly, even the participant most critical of prototype limitations selected the largest number of outputs for publication, underscoring the minimal sufficiency of the generated questions as exploratory entry points.

Participants also raised epistemic concerns regarding the situated and incomplete nature of the underlying KG, emphasising the importance of making data limits visible to end users. One historian noted that historical data is often fragmentary and context-dependent, and that a lack of explicit boundary-setting could lead users to over-interpret the apparent completeness of generated answers. This feedback highlights a design implication for Tús Maith beyond technical viability: transparency is required not only at the level of provenance and generation processes, but also at the level of corpus scope and representational limits.

A further concern raised by participants relates to the risk of prominence bias, whereby automated ranking and curation mechanisms may repeatedly surface the most well-represented entities or collections at the expense of less prominent but potentially significant ones. As one participant noted, similar dynamics in digitised archives have historically led to the marginalisation of less visible materials. This suggests that future instantiations of Tús Maith should consider incorporating diversity-aware or rotation-based strategies in entity ranking and CuQA surfacing to avoid reinforcing existing imbalances within the KG.

These findings support H1_UE3 while simultaneously highlighting the role of prototype fidelity in shaping expert evaluations of early-stage framework instantiations.

5.9.5 Summary of Key Findings

H1_UE3: "TMv1, an instantiation of the Tús Maith framework, is viable for domain experts to create CuQAs." is supported. Across participants, TMv1 enabled domain experts to efficiently produce filled-in questions and answers that were deemed suitable for publication as CuQAs and for use as exploratory entry points into the VRTI KG. Evidence for viability is demonstrated through consistent low perceived workload, moderate to high trust in system outputs, and the successful selection of publishable CuQAs by all participants, including those most critical of prototype limitations.

The extent to which H1_UE3 is supported can be understood through examination of the subsequent research questions that arose from the above hypothesis.

RQ1_UE3: "To what extent does TMv1 aid domain experts in the efficient creation of publishable CuQAs for the purpose of Tús Maith?"

TMv1 was found to meaningfully support domain experts in the creation of CuQAs intended as entry points for user exploration. All participants completed the TMv1 end-to-end within a single session and selected multiple filled-in questions for publication. The triangulated results indicate that TMv1 supports CuQA creation with minimal cognitive burden while maintaining sufficient output quality for expert acceptance. Variability in perceived suitability was primarily attributable to prototype expressiveness rather than to deficiencies in the underlying framework instantiation.

RQ1.1_UE3: "How usable and efficient do domain experts find TMv1?"

TMv1 was perceived as usable and efficient by domain experts. Modified RTLX scores indicated low to moderate workload, and task completion times demonstrated that CuQA curation could be performed efficiently, typically within approximately 11 minutes. While some participants spent longer in the final stage, this time was associated with reflective evaluation rather than confusion or frustration. These findings suggest that TMv1 is manageable both cognitively and temporally.

RQ1.2_UE3: "To what extent does the provision of provenance and model selection transparency provided by TMv1 promote trust in TMv1 within a domain expert?"

The provision of provenance and transparency features supported appropriate trust calibration rather than uniform reliance. Participants with lower baseline trust in AI actively engaged with provenance information for verification purposes, reporting that such fea-

tures were crucial for confidence in the outputs. Conversely, participants with higher trust interacted with provenance selectively or not at all, often citing curiosity rather than necessity. This pattern indicates that transparency mechanisms in TMv1 functioned as trust-supporting affordances, enabling scrutiny when needed without imposing mandatory overhead.

RQ1.3_UE3: "How do domain experts assess TMv1's suitability for supporting the creation of publishable CuQAs?"

Domain experts evaluated TMv1 as a suitable approach for generating CuQAs suitable for public-facing exploratory starting points within the constraints of the prototype. All participants selected at least two filled-in questions and answers for publication (creating CuQAs), with only one participant deciding to use the edit feature. Notably, the participant who expressed the strongest concerns regarding prototype limitations nevertheless selected the highest number of publishable CuQAs, indicating that perceived shortcomings related to expressive breadth rather than output correctness or usefulness. This supports the interpretation that TMv1 meets the minimal sufficiency requirements for CuQA generation, even under critical evaluation.

5.10 Assessing Tús Maith Framework's Portability: A Proof-of-Concept with LetterSampo

The following section presents a proof-of-concept using a separate, similarly sized historical KG to the VRTI KG. This proof-of-concept provides an initial test of whether Tús Maith, implemented as the TMv1 prototype, can be applied to a KG structurally different to the VRTI KG while remaining within the broader cultural heritage KG paradigm. A full evaluation of the extent of cross-KG applicability is beyond the scope of the present work and is therefore outlined as future work in Section 6.3.

5.10.1 Tús Maith Applied to the LetterSampo KG

To examine the portability of the Tús Maith framework beyond the VRTI KG, a proof-of-concept application was conducted using the LetterSampo correspondence KG [38], employing TMv1 as an instantiation of Tús Maith. A subset of nine CQ templates were selected from the full generated set aligning with LLM Experiment 2. Corresponding with the categories of information presented as exploratory perspectives in the LetterSampo UI, two of the most highly connected people, letters, and places were selected as instantiation entities in order to ensure sufficient semantic richness for CQ template instantiation.

The resulting filled-in questions and answers and automatic and manual verification of them indicate that TMv1 can operate over a structurally distinct humanities KG without modification to the underlying Tús Maith framework. As TMv1 constitutes a direct instantiation of Tús Maith, this provides some evidence that the Tús Maith framework itself is not dependent on any single ontology or dataset and can be applied across heterogeneous humanities KGs. This proof-of-concepts contributes additional supporting evidence toward Major Contribution 1.

The next subsections presents the details of the resulting proof-of-concept.

5.10.2 The Stages of Tús Maith as Applied to LetterSampo using TMv1

The LetterSampo KG contains approximately 2.9 million triples⁷ and is publicly accessible via the Linked Data Finland platform⁸. Like the VRTI ontology, it makes use of the CIDOC CRM ontology [12] to model people and events; however, it also incorporates its own custom Linked Data Finland (LDF) ontology, differentiating it structurally and semantically from VRTI.

GPT-5 was used for this instantiation, consistent with the TMv1 prototype, with temperature set to 1 and top_p set to 1. All code, LLM logs, and generated datasets are available in the GitHub repository⁹ associated with this thesis in the LetterSampo folder.

Stage 1: CQ Template Derivation

- **Ontology Extraction:** A generic SPARQL query (Appendix A.8.5.1) was used to retrieve classes and properties from the LetterSampo KG.
- **Ontology Compaction.** The ontology compaction process employed code nearly identical to that used for the VRTI KG, with only minor modifications arising from differences in how the ontology was exported (via direct endpoint querying rather than Virtuoso).
- **CQ Template Generation:** The compacted ontology was provided to GPT-5 using a prompt functionally identical to that used in the TMv1 prototype. The only substantive modifications were: (i) increasing the number of generated CQ templates from 20 to 40, and (ii) replacing references specific to the VRTI domain with correspondence-domain terminology. The full generated ontology, compacted ontology, and complete set of CQ templates are available in the associated GitHub repository folder.

In line with LLM Experiment 2 (Chapter 5) which utilised nine CQ templates, nine templates were selected from the generated set as can be seen in Listing 5.1. This selection is not intended as an evaluation of LLM template quality, but rather to maintain methodological consistency across TMv1 instantiations. Further, the selection of nine templates is not intended to provide representative coverage of all possible CQ template structures, but rather to verify that the template derivation and instantiation stages can be executed on a second KG. Specifically, the number and selection procedure align with: (i) LLM Experiment 1, where the number of generated templates scaled with ontology size (20 VRTI CQ Templates $\approx$ 18 VRTI KG Classes, 40 VRTI CQ Templates $\approx$ 41 LetterSampo KG Classes); (ii) User Evaluation 3, where a subset of templates was selected by the researcher; and (iii) LLM Experiment 2, where nine templates were used as illustrative examples.

⁷Number of Triples as of 08/02/2026 = 2,913,993

⁸<https://www.ldf.fi/dataset/corresp>, <https://seco.cs.aalto.fi/projects/rrl/>

⁹https://github.com/mcnamacl/T-s_Maith_Thesis_Repository

Listing 5.1: CQ Templates for LetterSampo

```
[
  "Which letters did a person write?",
  "To whom was a letter sent?",
  "From which place was a letter sent?",
  "Which ties involve a person?",
  "Which occupations does a person have?",
  "When was a person active?",
  "Which letters were sent from a place?",
  "Which external sources is a person linked to?",
  "How many correspondences does a person have?"
]
```

Stage 2: CQ Template Instantiation In the absence of usage analytics or access to a domain expert associated with the KG, connectivity-based prominence was used as a proxy for exploratory relevance, and alternative entity seeding strategies were employed.

1. The ten most highly connected person, place, and letter entities were retrieved using SPARQL, Appendix A.8.5.2.
2. From these, the two most highly connected entities of each type were selected. The semantic neighbourhood for each selected entity was generated using a generic SPARQL neighbourhood query (Appendix A.8.5.3).
3. The resulting semantic neighbourhoods were compressed using a script closely aligned with the ontology compaction procedure employed in Stage 1.

Filled-in questions and answers were then batch-generated by providing the LLM with the compressed semantic neighbourhoods of the six selected entities and the nine CQ templates produced in Stage 1.

Examples of the resulting filled-in questions are presented in Listing 5.2.

Listing 5.2: Examples of Filled-in Questions and Answers for LetterSampo produced by TMv1

```
{
  "results": [
    {
      "question": "Which letters did Haller, Albrecht von write?",
      "answer": "Haller, Albrecht von wrote 4,118 letters; examples include
        : 1728 Mar 7: Haller, Albrecht von (1708-1777) (Strassburg) to
        Burggrave, Johann Philipp (1700-1775) (Frankfurt am Main); 1728
        Sep 24: Haller, Albrecht von (1708-1777) (Bern) to Gessner,
        Johannes (1709-1790) (Z rich); 1729 Jun 19: Haller, Albrecht von
        (1708-1777) (Windisch) to Gessner, Johannes (1709-1790) (Basel).",
      "provenance": {
        "source_entity_uri": "http://ldf.fi/corresp/actors/
          p7395168262606659555"
      }
    }
  ]
}
```

```

},
{
  "question": "To whom was the letter '1520 Jul 22: Melanchthon,
    Philipp (Wittenberg) to Spalatin, Georg (Lochau)' sent?",
  "answer": "Spalatin, Georg.",
  "provenance": {
    "source_entity_uri": "http://ldf.fi/corresp/letters/lc1000"
  }
},
{
  "question": "From which place was the letter '1520 Jul 22:
    Melanchthon, Philipp (Wittenberg) to Spalatin, Georg (Lochau)'
    sent?",
  "answer": "Wittenberg.",
  "provenance": {
    "source_entity_uri": "http://ldf.fi/corresp/letters/lc1000"
  }
},
{
  "question": "Which ties involve Haller, Albrecht von?",
  "answer": "He is involved in 1,195 ties (actor2 in 1,093; actor1 in
    102). Examples: Achenwall, Gottfried (1719-1772) <-> Haller,
    Albrecht von; Adanson, Michel <-> Haller, Albrecht von; Haller,
    Albrecht von <-> Albinus, Bernhard Siegfried.",
  "provenance": {
    "source_entity_uri": "http://ldf.fi/corresp/actors/
    p7395168262606659555"
  }
},
{
  "question": "Which occupations does Melanchthon, Philipp have?",
  "answer": "Protestant reformer; astrologer; astronomer; classical
    philologist; evangelical theologian; humanist; philologist;
    philosopher; theologian; university teacher.",
  "provenance": {
    "source_entity_uri": "http://ldf.fi/corresp/actors/
    p2950305792108157270"
  }
},
{
  "question": "When was Haller, Albrecht von active?",
  "answer": "1728-1777.",
  "provenance": {
    "source_entity_uri": "http://ldf.fi/corresp/actors/
    p7395168262606659555"
  }
},
{
  "question": "Which letters were sent from Berlin?",
  "answer": "10,458 letters; examples include: 1540 Oct 4: Agricola,
    Johannes (Berlin) to Melanchthon, Philipp (Wittenberg); 1540 Oct
    4: Kf. Joachim (Berlin) to Melanchthon, Philipp (Wittenberg);

```

```

    1548 Aug 8: Buchholzer, Georg (Berlin) to Melanchthon, Philipp (
    Wittenberg).",
    "provenance": {
      "source_entity_uri": "http://ldf.fi/corresp/places/
        p1383671582106528001"
    }
  },
  {
    "question": "Which external sources is Melanchthon, Philipp linked to
      ?",
    "answer": "GND; EML0; DBpedia; ISNI; VIAF; Webpage; Wikipedia;
      Wikidata.",
    "provenance": {
      "source_entity_uri": "http://ldf.fi/corresp/actors/
        p2950305792108157270"
    }
  }
]
}

```

5.10.3 Validation Step

As there were no domain experts or GT available, to assess filled-in question and answer validity, the set of generated filled-in questions and answers were passed through the validation proxy presented in LLM Experiment 2, Section 5.7. The normalisation rules (see Appendix A.7) and checks were executed as per LLM Experiment 2, the code itself was only modified to account for the difference in source JSON structure. The full code and results can be found in LetterSampo GitHub folder.

More specifically, the three checks of the JSON validation proxy operated as before: Check 1 - ensure the source URI in the generated filled-in question and answer provenance exist in the source JSON; Check 2 - ensure after normalisation the value of the name in the filled-in question matches the value for the URI in the source JSON; Check 3 - ensure after normalisation that the filled-in answer matches any value in the source JSON. The results of validation proxy show that all generated filled-in questions and answers are valid as per the JSON validation proxy checks. The validity of filled-in questions and answers were further manually verified by the author of this thesis. This validation establishes factual grounding of generated outputs but does not constitute an assessment of their exploratory usefulness or domain relevance. The full results can be found in the LetterSampo folder of the thesis GitHub and a condensed example of the results is presented in Table 5.33.

5.10.4 LetterSampo Proof-of-Concept Conclusion

The LetterSampo proof-of-concept shows that the non-domain expert required stages (Stage 1 and Stage 2) of the Tús Maith framework, as instantiated in TMv1, can be executed on a second large-scale humanities KG without modification to the core framework. This provides preliminary evidence that the framework is not tightly coupled to the VRTI KG and can be applied to structurally distinct but semantically comparable cultural

Table 5.33: Validation Proxy Results.

Question	Answer	source_uri exists	entity_name matches	answer match_type	all_checks passed	is_exact match
Which ties involve Haller, Albrecht von?	He is involved in 1,195 ties (actor2 in 1,093; actor1 in 102). Examples: Achenwall, Gottfried (1719-1772) <-> Haller, Albrecht von; Adanson, Michel <-> Haller, Albrecht von; Haller, Albrecht von <-> Albinus, Bernhard Siegfried.	TRUE	TRUE	exact	TRUE	TRUE
From which place was the letter '1520 Jul 22: Melanchthon, Philipp (Wittenberg) to Spalatin, Georg (Lochau)' sent?	Wittenberg.	TRUE	TRUE	exact	TRUE	TRUE

heritage KGs.

However, this proof-of-concept represents a feasibility demonstration under favourable conditions, including the use of a CIDOC-CRM-aligned cultural heritage KG. It therefore establishes operational portability rather than full generalisability, which remains to be evaluated across a broader range of KG structures and domains.

5.11 Evaluation Conclusions

5.11.1 Cross-Evaluation Synthesis

This section synthesises findings across the five evaluations presented in this thesis by structuring the discussion around the two thesis research questions and their associated research objectives (RO3a - RO3c). It integrates evidence across human-centred and LLM-supported evaluations to assess the extent to which the Tús Maith framework can be considered as a coherent response to the Initial Exploration Problem (IEP). Collectively, these evaluations span the full CuQA lifecycle, from automated CQ template derivation and Instantiation (EX1-EX2), through domain expert curation (UE3), to in-situ use by lay users at the onset of KG exploration (UE1-UE2). The following sections present how the results of the evaluation contribute to answering the research questions of the thesis.

5.11.1.1 RQ1: Mitigation of the IEP for Lay Users

RQ1: "To what extent can domain expert curated natural language questions and answers (NLQAs) as entry points to RDF-based Knowledge Graphs, be perceived to mitigate the Initial Exploration Problem (IEP), as characterised in its framing, for lay users in the humanities domain?"

RQ1 was directly assessed via User Evaluation 1, Section 5.5 (UE1) and User Evaluation 2, Section 5.6 (UE2). UE1 suggested that, when compared to other available UI features, users with low prior familiarity with the VRTI Project preferred CuQAs as the feature to use to begin exploration. UE2 concentrated directly on CuQAs assessing the extent to which they were perceived to address the observable IEP barriers (scope uncertainty and ontology opacity) for lay users. The evaluation demonstrated that CuQAs addressed both barriers to large effect with statistical significance (when compared to neutral). Query incapacity was indirectly mitigated due to both CuQAs and the VRTI Explorer abstracting away from direct use of SPARQL.

It is therefore possible to conclude that CuQAs are perceived to address the observable barriers of the IEP as per its framing when compared to neutral for lay users beginning exploration of an unfamiliar, large, complex KG in the humanities domain and abstracts away the need for knowledge of formal query languages.

5.11.1.2 RQ2: Domain Expert Curation Support

RQ2: "To what extent does a multistage framework, when instantiated, support domain experts in the curation of such natural language question-answers (NLQAs) as entry points

to RDF-based Knowledge Graphs (KGs) in the humanities domain?"

User Evaluation 3, Section 5.9 (UE3) directly utilised the results of and lessons learned from LLM Experiment 1, Section 5.7, and LLM Experiment 2, Section 5.8, in order to address RQ2. Namely, CQ template derivation, Tús Maith Stage 1, when undertaken using LLMs, requires domain expert oversight, and CQ template instantiation, Tús Maith Stage 2, similarly when carried out using LLMs, requires domain expert review. The results of UE3 directly suggest that Tús Maith, instantiated as TMv1, is a viable means of generating CuQAs to be used as exploratory entry points for lay users. This establishes the instantiation of Tús Maith as a socio-technical system in which LLMs act as coverage amplifiers, while domain experts function as epistemic curators responsible for ensuring relevance, clarity, and answerability. Furthermore, the LetterSampo proof-of-concept presents some evidence of that Tús Maith framework, as instantiated creating TMv1, is generalisable beyond the VRTI KG.

Based on the evidence described in each of the evaluations presented in this chapter, Table 5.34 provides a summary of the extent to which the hypotheses directly relating to the research questions of this thesis and research objectives have been supported by the evidence obtained via the five evaluations. The nuance of what support vs. partial support means in the context of this thesis is presented in the next chapter.

Table 5.34: Research questions, hypotheses, and levels of support

Code	RQ	RO	Hypothesis	Level of Support
UE1	RQ1	RO3a	H1_UE1: The Tús Maith features provide new users with a useful starting point for beginning exploration of the VRTI KG.	Partially Supported
UE2	RQ1	RO3a	H1_UE2: CuQAs are perceived to address the observable barriers that characterise the IEP for lay users unfamiliar with the VRTI KG.	Partially Supported
EX1	RQ2	RO3b	H1_EX1: LLMs at their default settings (temperature=1, top_p=1) can generate relevant, clear, and answerable CQ templates suitable for Tús Maith when provided with information about the VRTI KG’s ontology. Suitable for Tús Maith here means that the CQs generated, when instantiated, could act as exploratory starting points for non-domain familiar lay users.	Partially Supported
	RQ2	RO3b	H2_EX1: LLMs can increase the coverage of a set of CQ templates for the VRTI KG (as judged by domain experts relative to an expert-authored ground truth (GT) set).	Supported
EX2	RQ2	RO3b	H1_EX2: LLMs at their default settings (temperature=1, top_p=1) can batch instantiate CQ templates to produce filled-in questions and answers using information about multiple entities. This was partially supported. The extent to which can be seen by looking at a more granular level via the experiment’s research questions.	Partially Supported
UE3	RQ2	RO3c	H1_UE3: TMv1, an instantiation of the Tús Maith framework, is viable for domain experts to create CuQAs.	Supported

Chapter 6

Conclusion

This final chapter of the thesis first presents the conclusions of the thesis; it then outlines the limitations of the current instantiation of Tús Maith and how such limitations translate into future directions for research in the area. The chapter next outlines the impact that this research is hoped to have on the wider academic community. Finally, it presents closing remarks from the author.

6.1 Research Objectives

This section discusses the extent to which each Research Objective (RO) has been addressed.

6.1.1 Research Objectives

6.1.1.1 RO1: To identify and characterise the barriers lay users face when beginning to explore unfamiliar complex KGs in the humanities domain (the barriers collectively characterise the Initial Exploration Problem (IEP)).

Addressed.

In identifying and conceptualising the Initial Exploration Problem (IEP) and its framing into three barriers grounded in human-computer interaction (HCI) literature, information behaviour theory, and KG usability literature, RO1 has been **addressed**.

Beyond identifying and characterising the difficulties associated with first contact of a KG by a lay user, this thesis analysed existing KG exploration interfaces through the lens of the IEP showing the primary assumption that they make – users already know where to begin. The emergence of solutions directly aimed at mitigating the IEP like OnSET, without having or providing a formal name for the problem, further highlights the need for a clear definition of barriers lay users face in order to develop more targeted interventions. Furthermore, this thesis examined KG exploration interface primitives, the epistemic preconditions required for such primitives to be effective, and where the composition of primitives cease to address the barriers of the IEP. It identifies scope revelation prior to user interaction as a missing interaction primitive that must be considered in order to mitigate

the IEP. Finally, in positioning the IEP and its framing as a design constraint as opposed to a usability flaw, this thesis has presented a means of articulating what constitutes an interface that addresses the IEP.

6.1.1.2 RO2: To identify and specify the lifecycle of NLQA entry points into KGs, and to design, instantiate, and implement a multistage framework over a complex digital humanities KG to support domain expert curation.

Addressed.

In identifying and specifying the lifecycle of Curated Questions and Answers (CuQAs), designing the three stage Tús Maith framework grounded in their lifecycle, instantiating the framework with support of large language models (LLMs) to produce TMv1, and implementing the instantiation over the VRTI KG, RO2 has been **addressed**.

The lifecycle of a CuQA is mapped to three stages of the Tús Maith framework in the following way: (1) Tús Maith Stage 1, CQ Template Derivation → Competency Question (CQ) Template; (2) Tús Maith Stage 2, CQ Template Instantiation → Filled-In Question and Answer; (3) Tús Maith Stage 3, Expert Curation → Curated Question and Answer (CuQA). Establishing this clear three stage curation process ensures that the CuQAs produced are grounded in the contents of the KG, are designed to expose the underlying structure of it via natural language (addressing *ontology opacity*), are, as a set, designed to convey the scope of the KG (addressing *scope uncertainty*), and are further designed to abstract away from query languages like SPARQL (addressing *query incapacity*). In adopting an entity-centric query model rather than direct CuQA → SPARQL translation, these CuQAs are designed to function as exploratory mechanisms presenting the answer to the curated question while also surfacing the semantic neighbourhood of entity/entities that form the basis for the answer; thus, distinguishing CuQAs from traditional NLQA paradigms. While instantiated in this thesis using the support of large language models (LLMs) and implemented over digital humanities KGs (the Virtual Record Treasury of Ireland (VRTI) KG as a use case and a proof-of-concept over the LetterSampo KG), in practice, as long as the responsibilities of the three stages are preserved, the framework is not tied to any particular method of instantiation or domain. However, it has yet to be determined whether it proves effective in alternate domains. LLMs as a method to support instantiation, resulting in TMv1, provide a scalable method of deriving and instantiating CQ templates to produce filled-in questions and answers the trade-off being increased domain expert oversight due to the stochastic nature of LLMs.

6.1.1.3 RO3a: Lay User Impact: To measure the extent to which curated NLQA entry points are perceived to mitigate the IEP, as per its framing, for lay users in the humanities domain.

Addressed.

The execution of both User Evaluation 1 (exploratory study on CuQAs in comparison to traditional exploratory interface features) and User Evaluation 2 (directly assessing the

extent to which CuQAs are perceived to address the observable barriers of the Initial Exploration Problem (IEP)) results in RO3a being **addressed**. User Evaluation 1, as an exploratory evaluation, indicated that CuQAs (when compared to two more traditional interface features) were seen as the most helpful exploration starting points feature by the majority of users unfamiliar with the KG. This led to User Evaluation 2 which concentrated on assessing the extent to which CuQAs alone were perceived to mitigate the observable barriers of the IEP. While User Evaluation 2 showed statistically significant improvements when compared to neutral on a 1-5 Likert Scale for lay users for two of the three barriers that characterise the IEP (scope uncertainty and ontology opacity), due to the nature of the implementation and evaluation, the third (query incapacity) was unable to be observed as users were not required to formulate queries. Furthermore, the construct of orientation, measured instead, did not show any measurable improvement. Taken together, these findings indicate that the hypothesis that CuQAs are perceived to mitigate the IEP is partially supported, in that mitigation was observed for some, but not all, IEP barriers. CuQAs were perceived to reduce scope uncertainty and improve understanding of the underlying structure of the KG, while evidence relating to query incapacity and broader orientation remains inconclusive within the scope of this thesis.

6.1.1.4 RO3b: To technically investigate and analyse Stage 1 and Stage 2 of the framework through its instantiation.

Addressed.

RO3b has been **addressed** through the technical investigation and analysis of Stage 1 and Stage 2 of the instantiation of the Tús Maith framework, TMv1, via LLM Experiment 1 and LLM Experiment 2. For the purposes of this thesis, the level of support assigned to each technical claim reflects the extent to which experimental evidence confirms that a given stage of the framework can be effectively realised through the chosen method of instantiation without requiring additional conditions.

LLM Experiment 1 established the design principle that remained a core component throughout the rest of the thesis: that LLMs can function as scaffolds but are not substitutes for domain experts. This result emerged from the fact that while those LLMs that were evaluated demonstrated the ability to produce CQ templates suitable for use within TMv1, and were able to increase CQ template coverage (generating CQ templates that could be added to the list of ground truth (GT) CQ templates created by the domain expert), there was not a uniform result. Some CQ templates produced by the LLMs were considered by domain experts to be not answerable by the KG, unclear, or not relevant as exploratory starting points. LLM Experiment 1 further highlighted the subjectivity of *relevance*, leading to its decomposition in User Evaluation 3. More concretely, LLM Experiment 1 suggests that LLMs can function in Stage 1 of TMv1 provided domain expert oversight is in effect, demonstrating that Stage 1 can be realised through LLM-based instantiation only under this condition.

LLM Experiment 2 further supported the above design principle while expanding on it to include interface-design considerations, namely the provision of provenance and model-

selection transparency. Specifically, the experiment demonstrated that the underlying LLM chosen greatly affects metrics such as runtime, consistency, accuracy, completeness, hallucination rate, and prompt compliance. As such, in line with LLM Experiment 1, while LLMs can be used as a method of instantiation of Tús Maith forming TMv1, this is only feasible as long as domain expert oversight remains integral to the process, demonstrating that Stage 2 can likewise be realised through LLM-based instantiation only under these conditions.

Thus, across both experiments, the claim that LLMs can fulfil the responsibilities of Stage 1 and Stage 2 is **partially supported**, as domain expert intervention during Stage 1 and Stage 2 is non-optional. The claim that LLMs increase coverage of CQ templates, however, was supported. Taken together, the technical investigations demonstrate that LLMs, when coupled with domain expert oversight, can function effectively within Stage 1 and Stage 2 of TMv1. In establishing the technical behaviour, feasibility, and constraints of these stages under LLM-based instantiation, RO3b has been fully addressed.

6.1.1.5 RO3c: Domain Expert Evaluation: Evaluate with domain experts the viability of the instantiated framework.

Addressed.

RO3c has been **addressed** through User Evaluation 3, which directly evaluated the viability of the instantiated framework with domain experts.

Through the incorporation of outcomes from both LLM experiments (prompt design, entity ranking strategies, specific LLM to be used, provenance and model selection transparency), a prototype TMv1 system was implemented over the VRTI KG. This prototype constituted a direct implementation of the design constraints and insights derived from LLM Experiment 1 and LLM Experiment 2. Whereas the earlier experiments examined the behaviour of LLMs in Stage 1 and Stage 2 in isolation, TMv1 represents their integrated execution within a single operational system. As such, User Evaluation 3 examined the stages, not as discrete technical components, but as a cumulative domain expert curation system. In addition, the evaluation was intentionally scoped as a controlled end-to-end demonstration of the instantiated framework in practice, with system components (Stage 1 CQ templates, and Stage 2 entities) configured in advance based on the outcomes of the previous LLM Experiments. More precisely, User Evaluation 3 assessed TMv1’s viability as an instantiation of Tús Maith by decomposing "viable" into three measurable constructs: system trust, TMv1 suitability, and workload. Across all constructs, the results of the evaluation suggest that TMv1, as an instantiation of Tús Maith, can function as a system to curate CuQAs, and as such the claim that it is viable is supported.

The findings indicate that domain experts were able to use TMv1 to carry out the curation process with an acceptable level of system trust, perceived suitability, and workload, demonstrating that the instantiated framework can be utilised in practice as a domain expert curation system. Further investigation into long-term use by domain experts was out of scope for this thesis. As will be discussed in the Limitations section below, assessing TMv1’s in its current state (namely, its ability to generate simple CuQAs) was seen

as a necessary pre-cursor to the implementation and examination of it to create complex CuQAs.

In establishing the practical viability of the instantiated framework for domain expert use, RO3c has been addressed.

6.1.2 Research Contributions

In Chapter 1, one major and two minor contributions of this thesis were presented. These contributions are discussed further below. All empirical findings in this thesis are derived from implementations within the digital humanities domain; while the IEP and CuQA lifecycle are theorised as domain-agnostic, cross-domain validation remains future work.

6.1.3 Major Contribution: The Tús Maith Framework for Curated Natural Language Exploratory Entry Points into Knowledge Graphs (KGs)

The primary contribution of this thesis is the identification and specification of the lifecycle through which Curated Natural Language Question and Answer (CuQA) entry points into KGs are produced, and the design, instantiation, implementation, and evaluation of Tús Maith, a multistage framework grounded in this lifecycle. Tús Maith establishes a structured framework through which domain experts can create CuQAs that act as exploratory starting points for lay users in mitigation of the Initial Exploration Problem (IEP).

The research conducted for this thesis demonstrates that CuQAs can operate as cognitive and linguistic scaffolds that reveal both the scope and structure of a KG without requiring prior knowledge of its ontology, a formal query language, or its contents. The framework structures the lifecycle into three stages: Competency Question (CQ) template derivation, CQ template instantiation, and expert curation. In repurposing CQs from ontology engineering artefacts into exploratory instruments, this thesis establishes a practical mechanism for exposing KG contents through natural language. Furthermore, in decoupling the CuQAs directly from SPARQL but instead surfacing the instantiating entity or entities' semantic neighbourhood in answering the question, the thesis presents an entity-centric query model that enables NLQAs to function as exploratory mechanisms as opposed to information retrieval artefacts.

Empirical evaluation shows that such CuQAs are perceived by lay users to reduce scope uncertainty and improve understanding of a KG's structure providing partial support for their role in mitigating the IEP. Further evaluation demonstrates that the instantiated framework is viable for domain experts to curate such entry points in practice. Together, these findings establish Tús Maith as both a conceptual and practical contribution: it demonstrates that the onset of KG exploration can be supported through curated NLQA (CuQA) entry points and provides an implementable framework for producing them.

6.1.4 Minor Contribution 1: The Identification and Characterisation of the Initial Exploration Problem

This thesis identifies and characterises the Initial Exploration Problem (IEP), defined as the set of three interrelated barriers encountered by lay users attempting to begin exploration of an unfamiliar complex KG: **scope uncertainty**, **ontology opacity**, and **query incapacity**. Grounded in human-computer interaction (HCI), information behaviour, and KG usability literature, this framing provides the conceptual basis for analysing first-contact KG experiences and for designing interventions that target the onset of exploration.

In applying this framing to existing KG exploration interfaces, this thesis shows that many assume prior user knowledge of where to begin. In response, it positions scope revelation prior to interaction as a missing interaction primitive and frames the IEP as a design constraint for KG interfaces rather than solely a usability deficiency. This contribution provides a conceptual vocabulary and design-oriented framing for further work on KG accessibility and exploration.

6.1.5 Minor Contribution 2: Empirical Evaluations of large language model (LLM)-Assisted Scaffolding for KGs

A further contribution of this thesis lies in the empirical investigation of how large language models (LLMs) can function as scaffolds within the instantiation of the Tús Maith framework (TMv1) to support domain experts in the creation of Curated Natural Language Questions and Answers (CuQAs). Within this research, domain experts responsible for curating CuQAs as exploratory entry points are assumed to possess deep subject knowledge but may lack technical familiarity with the underlying KG structure and formal query languages. As such, they themselves may encounter barriers of the IEP, particularly ontology opacity and query incapacity. The use of LLMs within TMv1 is therefore positioned not as a replacement for domain expert expertise, but as a supportive mechanism that assists them in deriving and instantiating CQ templates grounded in the KGs content.

Empirical investigation across LLM Experiment 1 and LLM Experiment 2 demonstrate that LLMs can assist in expanding the breadth of potential CQ templates, generate answerable, clear, and relevant CQ templates, and instantiate CQ templates with structured entity information. However, the findings also show that LLM outputs are not uniformly suitable for direct use and as such, require sustained domain expert oversight. These results establish that LLMs within TMv1 function most effectively as scaffolds that support rather than replace domain experts.

The research further surfaces design considerations for integrating LLMs into domain expert curation systems, in particular, the provision of provenance information and model-selection transparency in order to aid in LLM-explainability, traceability, and function as trust-calibration affordances for domain experts. These considerations define practical constraints for the responsible integration of LLMs into structured KG-based systems.

A further methodological contribution arises from the design and evaluation of a lightweight JSON-based validation proxy for assessing LLM-generated CQ template instantiation outputs against source-entity data in the absence of GT. The results show that such structured

validation can function as an effective first-pass filter within curator-facing systems like TMv1 enabling rapid identification of structurally valid filled-in questions and answers. While not a replacement for human validation, it establishes a practical mechanism for integrating automated validation into LLM-supported KG curation systems.

Taken together, this contribution provides an empirically grounded account of how LLM-assisted scaffolding and lightweight validation mechanisms can support domain expert curation of CuQAs as exploratory entry points within the context of complex digital humanities KGs, while also identifying the practical constraints and design requirements associated with such integration.

The next section presents the limitations of claims and TMv1 as it was implemented in this thesis, clarifying the scope within which the contributions should be considered.

6.2 Limitations

The findings of this thesis should be interpreted within the methodological and empirical scope of the evaluations conducted, synthesis of existing literature, and the implementation of the Tús Maith framework through its instantiation, TMv1.

6.2.1 Limitations of Claims

Firstly, the Initial Exploration Problem (IEP) is presented in this thesis as a conceptual framing synthesised from human-computer interaction (HCI), information behaviour, and KG usability literature. The three interrelated barrier composition provides a lens with which to examine first-contact KG exploration and for designing targeted interventions. While the IEP is theorised as domain-agnostic, and therefore has been used to analyse existing KG exploration primitives and exploration interfaces, investigation of CuQAs as the empirical evaluation work presented as part of this thesis is contained within the context of a complex digital humanities KG. Further work across additional domains and KG structures therefore would strengthen understanding of how the manifestation of individual barriers may vary. This does not undermine the conceptual contribution of the IEP, but rather defines the empirical scope within which it has been examined in this work.

Secondly, the CuQA lay user evaluations (User Evaluation 1 and User Evaluation 2) relied primarily on convenience sampling with participants drawn largely from third-level students with an implicit interest in history and cultural heritage. While this ensured meaningful engagement with the VRTI KG, it restricts generalisability. The findings therefore support claims about unfamiliar but domain-interested users, not about all lay users. Behaviour and perception may differ in non-academic populations, in other age-groups, or in lay users without intrinsic motivation to explore cultural heritage data.

6.2.2 Limitations of TMv1

TMv1 represents an early instantiation of the Tús Maith framework and demonstrates its viability in supporting domain experts in the creation of curated questions and answers (CuQAs), as evidenced by User Evaluation 3 in Chapter 5. However, as a research pro-

tototype, it inherits constraints arising from its initial design choices (supporting curation of only simple CuQAs), LLM-supported implementation, and single-domain deployment. These limitations define the scope within which the present findings should be interpreted. One limitation is that TMv1 currently produces primarily single-hop CuQAs that map to properties of individual entities without supporting multi-hop semantic traversals. Feedback from domain experts in User Evaluation 3 indicated a desire for semantically richer, multi-hop questions capable of revealing deeper relational structure within the KG. The current focus on CuQAs that can be filled in by a single entity’s semantic neighbourhood reflected a deliberate decision prioritising reproducibility, validation, and oversight rather than a restriction imposed by the Tús Maith framework itself. More specifically, the research presented here concentrated on investigating whether a three-stage framework (Tús Maith) to curate NLQA KG entry-points (CuQAs), when instantiated using LLMs (TMv1) had the capacity to support domain experts in CuQA creation; treating its ability to produce simple CuQAs as a necessary pre-cursor to the production of complex CuQAs. While sufficient to examine the viability of TMv1 by domain experts, this limits the expressive depth of curated entry points. Future extensions to TMv1 might for example incorporate controlled multi-hop traversal through structured prompting with graph-based traversal modules, as outlined in Appendix B.

Second, TMv1’s reliance on LLMs in Stages 1 and 2 introduces variability associated with stochastic generation. Although GPT-5 demonstrated strong performance in LLM Experiment 2 (Chapter 5), including low hallucination rates under the conditions tested, output quality remains sensitive to prompt design, model configuration, and versioning. The use of LLMs also introduces trade-offs in cost, latency, and environmental impact. These factors necessitate a conservative human-in-the-loop governance model in which domain experts retain final decision authority over all curated outputs. While this approach supports epistemic reliability and interpretive validity, it increases verification effort and limits the degree to which CuQA production can be fully automated.

Third, the mandatory role of expert curation introduces a scalability constraint. Human judgement is essential to ensure that curated entry points are relevant, clear, and epistemically appropriate, particularly in interpretive domains such as the digital humanities. However, this requirement bounds throughput and means that the rate of CuQA production is contingent on curator availability. TMv1 is therefore designed to reduce, rather than eliminate, expert labour in the creation of exploratory entry points.

Fourth, TMv1 has been deployed and evaluated using a single KG, the Virtual Record Treasury of Ireland (VRTI) KG. While the VRTI KG provides a complex and highly suitable test case due to its scale, heterogeneity, and intended audience including lay users, reliance on a single deployment context limits empirical generalisability. The Tús Maith framework itself is designed to be domain- and KG-agnostic, requiring only access to schema information, entity materialisation mechanisms, and optional ranking signals. Nonetheless, further instantiations across additional domains and KGs will be required to fully assess transferability. A proof-of-concept demonstrating the portability of Tús Maith beyond the VRTI KG, implemented on an additional KG, is presented in Chapter 5.10; however, broader

multi-domain evaluation remains an important direction for future work.

Finally, as a prototype system evaluated under controlled conditions, certain components of TMv1 were fixed or precomputed for the purposes of User Evaluation 3. While this supported internal validity and ensured consistency across participants, it limits conclusions about fully dynamic deployment at scale. These constraints reflect the exploratory nature of the present instantiation rather than limitations of the underlying framework.

Taken together, these limitations define the operational boundaries of TMv1 as implemented and evaluated in this thesis, while also indicating clear pathways for extension and cross-domain validation.

6.3 Future Work

Future work examines two primary limitations of the current implementation of TMv1. The first is that it has only been evaluated using one KG, the VRTI KG. The second is the lack of explicit support for the generation of multi-hop CuQAs. Future work addressing both of these limitations is outlined below.

6.3.1 Applying TMv1 beyond the VRTI KG and Proof-of-Concept

As discussed in the previous sections, the evaluation of this research was situated within the digital humanities domain. Specifically, all five evaluations were carried out using the VRTI KG. While this provided a rich test-bed with which to examine the IEP, CuQAs, and the TMv1, it limits the examination generalisability. Through the proof-of-concept with the LetterSampo KG, Stage 1 and Stage 2 of TMv1 have been shown, with minimal edits, to produce valid filled-in questions and answers (as per the JSON validation proxy and manual-author verification) over a different digital humanities KG. Future work should not only investigate whether the filled-in questions and answers produced are considered viable as entry points into the LetterSampo KG, in other words, are considered by domain experts of LetterSampo to be CuQAs, but also, conduct evaluations with TMv1 (or similar) over KGs in completely distinct domains.

6.3.2 Support for Multi-Hop, Complex CuQA Creation

Domain expert feedback obtained during the evaluation of TMv1 indicated a desire for CuQAs that move beyond single-relation or attribute-centric frames toward questions that capture broader relational context. Future work therefore should investigate the computational feasibility of enabling this. Appendix B presents a potential approach to implementing multi-hop CuQA curation suggesting the coupling of graph topology with meta-paths to achieve this. Another approach could be the inclusion of richer ontological information obtained from the ontologies utilised by the KG when deriving CQs from the KG.

6.3.3 Support for Multilingualism

The research conducted as part of this thesis focused on CuQA curation in English. An interesting future direction could be to examine the impact of creating CuQAs using the

Tús Maith framework in other languages (e.g., Spanish) from KGs that may or may not be in the language in question. In doing so, questions are raised surrounding CuQA generation and translation, as well as ensuring that the CuQA remains adequately grounded in the KG’s structure, vocabulary, and content.

6.4 Impact

It is hoped that the research undertaken as part of this thesis will contribute to new research directions to be explored and will shine a greater spotlight on the moment of first contact between a lay user and a KG. This section provides some of the ways that this may occur.

6.4.1 Research Dissemination

Almost all of the research in this thesis has been disseminated either through peer-reviewed publications (LLM Experiment 1 at the First International Workshop on Users and Knowledge Graphs @ SEMANTiCS 2025, LLM Experiment 2 at the ACM Intelligent User Interfaces conference 2026), submission to a relevant journal (User Evaluation 2 and User Evaluation 3 submitted to the Journal on Computing and Cultural Heritage 2026), or as a preprint (the conceptualisation of the Initial Exploration Problem and where it sits within literature as a preprint on arXiv). Such dissemination allows for potential uptake within the semantic web community and beyond of CuQAs as scaffolds for entry, the Initial Exploration Problem (IEP) as vocabulary to describe the barriers at first contact, and the Tús Maith framework as a framework to curate CuQAs. The rest of this section suggests how the work could impact the wider research community.

6.4.2 Potential Impact

In giving language and structure to barriers hindering lay users from beginning exploration of a large complex KG through the conceptualisation of the Initial Exploration Problem (IEP), this thesis has introduced targets to address for interfaces aiming to support these lay users. It provides a unified lens with which to compare existing and build new interaction primitives and combinations of such potentially enabling the creation of better overall solutions focused on supporting initial lay user exploration.

Tús Maith, as a framework for domain experts to curate CuQAs, puts forward a number of new ideas that have the potential to inspire new lines of inquiry. Firstly, repurposing CQs to form reusable templates supporting exploration, in contrast to their original purpose of aiding ontology engineers to assess KG completeness, provides a new way of considering CQs. Such re-framing of their purpose may lead to other avenues beyond CuQAs being explored e.g., using CQs in information retrieval NLQA contexts. Secondly, adopting an entity-centric query model allows for NLQAs to function not just as information retrieval mechanisms but as exploration enablers. Once again, this could branch out into interesting research avenues combining recommendation systems ("if you looked at that, you may be interested in this") with the semantic neighbourhood(s) of the entity/entities returned.

Thirdly, considering NLQAs not just as usability features but as epistemic instruments shaping users' mental models of the underlying KG could lead to interesting research looking more deeply into how this connects with sensemaking [29, 166] and information behaviour theories such as ASK [65].

While CuQAs are not currently implemented within the public facing VRTI Explorer, they have been implemented within the code of the VRTI Explorer as Featured Questions as part of User Evaluation 2. This leaves potential for them to be considered as an addition to the VRTI Explorer, especially as User Evaluation 3 indicated that Tús Maith, instantiated as TMv1, was deemed viable to curate simple CuQAs by domain experts. In other words, an implementation exists, Featured Questions showed promise from User Evaluation 2 and the framework to create them, Tús Maith, as TMv1, further showed promise in assisting domain experts curate them.

The overarching outcomes of the technical experiments with LLMs, LLM Experiment 1 and LLM Experiment 2 have resulted in the design principle "LLMs as scaffolds not substitutes for domain experts". Extrapolating this from the empirical evidence and seeing it reflected across both experiments and into User Evaluation 3 with the TMv1 prototype, will hopefully further discourse on the place of AI tools such as LLMs within contexts such as the digital humanities.

Finally, it is hoped that overall, this research will make KGs more accessible by communities outside of the semantic web community potentially leading to greater uptake and continued use of KGs in the digital humanities and beyond.

6.5 Final Remarks

The problem of "where to start" is not one that is unique to the semantic web. Nor even, is it unique to academic domains. It is one encountered often across all areas of life from the "where to start" when staring at a blank page, to a mound of washing that has yet to be tackled. What is different here however, within KGs, is the often complex nature of them compounded by scale and unfamiliarity. Furthermore, within the context of exploration, the absence of a clearly defined goal necessitates the examination of the problem not just through the lens of usability, but instead also through epistemology. In shifting perspective from one of exploration optimisation to instead one of mental model construction, the burden of such construction is placed on the interfaces and features that make them up rather than on the user. The research here argues that such a shift is imperative if the barriers of the Initial Exploration Problem (IEP) are to be addressed and the conceptualisation of the IEP enables this through shared language in clearly defining these barriers. Artefacts such as Curated Question and Answers (CuQAs) can therefore be analysed from the dual viewpoints of usability and epistemology and Tús Maith as a framework to curate them can be examined in kind. It is hoped that considering the problem of "where to start" from a multi-disciplinary perspective and the introduction of mental scaffolds like CuQAs curated via Tús Maith will bring a fresh perspective of KG exploration as after all, *Tús Maith Leath na hOibre*, a good start is half the work!

Generative AI Disclosure

During the preparation of this thesis, the author used generative AI tools (large language models) in a limited assistive capacity. These tools supported tasks such as proofreading draft text, improving clarity of phrasing, discussing possible ways of structuring sections of the manuscript, and scaffolding or debugging code used in the research. Any such code was reviewed, tested, and modified by the author, with further documentation provided in the associated repository READMEs. These tools were not used to generate original research contributions, analyses, or conclusions. The author takes full responsibility for the final content of the thesis.

Bibliography

- [1] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D’amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. Knowledge graphs. *ACM Comput. Surv.*, 54(4), July 2021. ISSN 0360-0300. doi: 10.1145/3447772. URL <https://doi.org/10.1145/3447772>.
- [2] RDF. Resource description framework (rdf). <http://www.w3.org/RDF/>, 2014. [Online; Accessed June 2025].
- [3] David Beckett. RDF 1.1 Turtle (Terse RDF Triple Language). W3C Recommendation, February 2014. URL <https://www.w3.org/TR/turtle/>.
- [4] Kris McGlinn, Matthew A Rutherford, Karl Gisslander, Lucy Hederman, Mark A Little, and Declan O’Sullivan. Fairvasc: A semantic web approach to rare disease registry integration. *Computers in Biology and Medicine*, 145:105313, 2022.
- [5] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The Semantic Web (ISWC 2007)*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer, 2007. doi: 10.1007/978-3-540-76298-0_52. URL https://link.springer.com/chapter/10.1007/978-3-540-76298-0_52.
- [6] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. Dbpedia – a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195, 2015. doi: 10.3233/SW-140134. URL <https://journals.sagepub.com/doi/abs/10.3233/SW-140134>.
- [7] Wikimedia Foundation. Wikipedia, the free encyclopedia. <https://www.wikipedia.org/>, 2001. URL <https://www.wikipedia.org/>. [Online; Accessed November 2025].
- [8] Aba-Sah Dadzie and Matthew Rowe. Approaches to visualising Linked Data: A survey. *Semantic Web*, 2(2):89–124, 2011. ISSN 15700844. doi: 10.3233/SW-2011-0037. URL <https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/SW-2011-0037>.
- [9] Matthew K Gold. *Debates in the digital humanities*. U of Minnesota Press, 2012.

- [10] Virtual Record Treasury of Ireland. Knowledge graph. Online resource, 2025. URL <https://virtualtreasury.ie/knowledge-graph>. Online; Accessed June 2025.
- [11] Virtual Record Treasury of Ireland. Virtual Record Treasury of Ireland. Online resource, 2025. URL <https://virtualtreasury.ie/>. Online; Accessed June 2025.
- [12] Martin Doerr. The cidoc conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI Mag.*, 24(3):75–92, September 2003. ISSN 0738-4602.
- [13] Robert Battle and Dave Kolas. Geosparql: enabling a geospatial semantic web. *Semantic Web Journal*, 3(4):355–370, 2011.
- [14] Lucy McKenna, Lynn Kilgallon, Alex Randles, Beyza Yaman, Christophe Debruyne, Fabrizio Orlandi, Gary Munnelly, Peter Crooks, and Declan O’Sullivan. Virtual Record Treasury of Ireland (VRTI) Ontology (v1.2). Web page, January 2025. URL <http://ont.virtualtreasury.ie/ontology/index-en.html>. Revision v1.2, Accessed June 2025.
- [15] H. Li, G. Appleby, C. D. Brumar, R. Chang, and A. Suh. Knowledge graphs in practice: Characterizing their users, challenges, and visualization opportunities. *arXiv preprint*, arXiv:2304.01311, 2023.
- [16] SPARQL. Sparql query language for rdf. <https://www.w3.org/TR/sparql11-query/>, 2013. [Online; Accessed June 2025].
- [17] Matteo Lissandrini, Torben Bach Pedersen, Katja Hose, and Davide Mottin. Knowledge graph exploration: where are we and where are we going? *SIGWEB Newsl.*, 2020(Summer), July 2020. ISSN 1931-1745. doi: 10.1145/3409481.3409485. URL <https://doi.org/10.1145/3409481.3409485>.
- [18] Guillermo Vega-gorgojo, Laura Slaughter, Bjørn Marius Zernichow, Nikolay Nikolov, and Dumitru Roman. Linked data exploration with rdf surveyor. *IEEE Access*, PP: 1–1, 11 2019. doi: 10.1109/ACCESS.2019.2956345.
- [19] Emil Kuric, Javier Fernández, and Olha Drozd. *Knowledge Graph Exploration: A Usability Evaluation of Query Builders for Laypeople*, pages 326–342. 11 2019. ISBN 978-3-030-33219-8. doi: 10.1007/978-3-030-33220-4_24.
- [20] Sampo ui. <https://seco.cs.aalto.fi/tools/sampo-ui/>, n.d. [Online; Accessed June 2025].
- [21] Martins Zviedris and Guntis Barzdins. Viziquer: a tool to explore and query sparql endpoints. In *Extended Semantic Web Conference*, pages 441–445. Springer, 2011.
- [22] Jiho Kim, Yeonsu Kwon, Yohan Jo, and Edward Choi. KG-GPT: A General Framework for Reasoning on Knowledge Graphs Using Large Language Models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9410–9421, Singapore, December 2023.

- Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.631. URL <https://aclanthology.org/2023.findings-emnlp.631/>.
- [23] Vincent Emonet, Jerven Bolleman, Severine Duvaud, Tarcisio Mendes de Farias, and Ana Claudia Sima. Llm-based sparql query generation from natural language over federated knowledge graphs. *arXiv preprint arXiv:2410.06062*, 2024.
 - [24] Rungsiman Nararatwong, Natthawut Kertkeidkachorn, and Ryutaro Ichise. Knowledge graph visualization: Challenges, framework, and implementation. In *2020 IEEE Third International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 174–178. IEEE, 2020.
 - [25] World Historical Gazetteer, n.d. URL <https://docs.whgazetteer.org/content/v4/data-model/rdf-representation.html#complete-example>. [Online; Accessed November 2025].
 - [26] Paul Resnick and Hal R. Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, March 1997. ISSN 0001-0782, 1557-7317. doi: 10.1145/245108.245121. URL <https://dl.acm.org/doi/10.1145/245108.245121>.
 - [27] Hongli Yuan and Alexander Hernandez. User cold start problem in recommendation systems: A systematic review. *IEEE Access*, 12 2023. doi: 10.1109/ACCESS.2023.3338705.
 - [28] Benedikt Kantz, Kevin Innerebner, Peter Waldert, Stefan Lengauer, Elisabeth Lex, and Tobias Schreck. Onset: Ontology and semantic exploration toolkit, 2025. URL <https://arxiv.org/abs/2504.08373>.
 - [29] Daniel M. Russell, Mark J. Stefik, Peter Pirolli, and Stuart K. Card. The cost structure of sensemaking. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, CHI '93, page 269–276, New York, NY, USA, 1993. Association for Computing Machinery. ISBN 0897915755. doi: 10.1145/169059.169209. URL <https://doi.org/10.1145/169059.169209>.
 - [30] Sébastien Ferré. SPARKLIS: An Expressive Query Builder for SPARQL Endpoints with Guidance in Natural Language. *Open Journal Of Semantic Web*, 0, 2017. URL <https://inria.hal.science/hal-01485093>.
 - [31] Silvio Peroni, David Shotton, and Fabio Vitali. Tools for the automatic generation of ontology documentation: A task-based evaluation. *Int. J. Semant. Web Inf. Syst.*, 9(1):21–44, January 2013. ISSN 1552-6283. doi: 10.4018/jswis.2013010102. URL <https://doi.org/10.4018/jswis.2013010102>.
 - [32] Sean M. Falconer and Margaret-Anne Storey. A Cognitive Support Framework for Ontology Mapping. In Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *The Semantic*

- Web*, pages 114–127, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-76298-0.
- [33] Karl Grossner and Ruth Mostern. Linked places in world historical gazetteer. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Geospatial Humanities*, GeoHumanities '21, page 40–43, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450391023. doi: 10.1145/3486187.3490203. URL <https://doi.org/10.1145/3486187.3490203>.
 - [34] Michael Grüninger and Mark Fox. The role of competency questions in enterprise engineering. *IFIP Advances in Information and Communication Technology*, 03 1995. doi: 10.1007/978-0-387-34847-6_3.
 - [35] Glaice Kelly Q. Monfardini, Jordana S. Salamon, and Monalessa P. Barcellos. Use of competency questions in ontology engineering: A survey. In João Paulo A. Almeida, José Borbinha, Giancarlo Guizzardi, Sebastian Link, and Jelena Zdravkovic, editors, *Conceptual Modeling*, pages 45–64, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-47262-6.
 - [36] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99:101805, 2023. ISSN 1566-2535. doi: <https://doi.org/10.1016/j.inffus.2023.101805>. URL <https://www.sciencedirect.com/science/article/pii/S1566253523001148>.
 - [37] Alex Randles, Lucy McKenna, Lynn Kilgallon, Beyza Yaman, Peter Crooks, and Declan O’Sullivan. The knowledge graph explorer for the virtual record treasury of ireland. In *Proceedings of the 9th International Workshop on Visualization and Interaction for Ontologies, Linked Data and Knowledge Graphs (VOILA 2024)*, Baltimore, USA, 2024. Co-located with the 23rd International Semantic Web Conference (ISWC 2024), November 11–15, 2024.
 - [38] Eero Hyvönen, Petri Leskinen, and Jouni Tuominen. Lettersampo—historical letters on the semantic web: A framework and its application to publishing and using epistolary data. *ACM Journal on Computing and Cultural Heritage*, 16(1):1–23, 2023.
 - [39] James R Lewis. Psychometric evaluation of the post-study system usability questionnaire: The pssuq. In *Proceedings of the human factors society annual meeting*, volume 36, pages 1259–1260. Sage Publications Sage CA: Los Angeles, CA, 1992.
 - [40] John Brooke. Sus: A quick and dirty usability scale. *Usability Eval. Ind.*, 189, 11 1995.

- [41] Fred D Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 13(3):319–340, 1989.
- [42] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL <https://arxiv.org/abs/1908.10084>.
- [43] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT, February 2020. URL <http://arxiv.org/abs/1904.09675>. arXiv:1904.09675 [cs].
- [44] Claire McNamara, Lucy Hederman, and Declan O’Sullivan. Mind the gap: LLMs, competency questions, and the non-technical user in the humanities domain. In *First International Workshop on Users and Knowledge Graphs @ SEMANTiCS 2025*, 2025. URL <https://openreview.net/forum?id=gDzqvLCrgc>.
- [45] Claire McNamara, Lucy Hederman, and Declan O’Sullivan. Filling the gap: LLMs as scaffolds for competency question instantiation. In *Proceedings of the ACM Intelligent User Interfaces ’26 Conference*, 2026. ISBN 979-8-4007-1984-4/26/03. doi: 10.1145/3742413.3789141. URL <https://doi.org/10.1145/3742413.3789141>.
- [46] Claire McNamara, Lucy Hederman, and Declan O’Sullivan. The initial exploration problem in knowledge graph exploration, 2026. URL <https://arxiv.org/abs/2602.21066>.
- [47] Anna Sofia Lippolis, Minh Davide Ragagni, Paolo Ciancarini, Andrea Giovanni Nuzozolese, and Valentina Presutti. Bench4ke: Benchmarking automated competency question generation, 2025. URL <https://arxiv.org/abs/2505.24554>.
- [48] Fiorela Ciroku, Jacopo de Berardinis, Jongmo Kim, Albert Meroño-Peñuela, Valentina Presutti, and Elena Simperl. Revont: Reverse engineering of competency questions from knowledge graphs via language models. *Journal of Web Semantics*, 82:100822, 2024. ISSN 1570-8268. doi: <https://doi.org/10.1016/j.websem.2024.100822>. URL <https://www.sciencedirect.com/science/article/pii/S1570826824000088>.
- [49] Reham Alharbi, Valentina Tamma, Floriana Grasso, and Terry Payne. An experiment in retrofitting competency questions for existing ontologies. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing, SAC ’24*, page 1650–1658. ACM, April 2024. doi: 10.1145/3605098.3636053. URL <http://dx.doi.org/10.1145/3605098.3636053>.
- [50] Reham Alharbi, Jacopo de Berardinis, Terry R Payne, and Valentina Tamma. Comparing methods for competency question elicitation from ontology requirements. 2025.
- [51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

- [52] OpenAI. Gpt-4o, 2024. URL <https://platform.openai.com/docs/models/gpt-4o>. Online; Accessed June 2025.
- [53] OpenAI. Openai gpt-5 system card, 2025. URL <https://arxiv.org/abs/2601.03267>.
- [54] GoogleCloud. Gemini 2.0 flash, 2024. URL <https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash>. Online; Accessed June 2025.
- [55] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL <https://arxiv.org/abs/1907.11692>.
- [56] Beyza Yaman, Alex Randles, Lucy McKenna, Lynn Kilgallon, Diego Rincón-Yáñez, Neil Johnston, Peter Crooks, and Declan O’Sullivan. Expanding the virtual record treasury of ireland knowledge graph. *Semant. Web J.*, 2024.
- [57] Peter Crooks, Zoë Reid, and Ciarán Wallace. The virtual record treasury of ireland: A century of recovery from the 1922 four courts blaze—and beyond. *History Ireland*, 30(3):pp. 38–41, 2022. ISSN 07918224. URL <https://www.jstor.org/stable/27198424>.
- [58] Orri Erling and Ivan Mikhailov. Rdf support in the virtuoso dbms. In *Networked Knowledge-Networked Media: Integrating Knowledge Management, New Media Technologies and Semantic Systems*, pages 7–24. Springer, 2009.
- [59] Lodview. <https://www.w3.org/wiki/LodView>, 2014. [Online; Accessed June 2025].
- [60] Dmitry Mouromtsev, Dmitry Pavlov, Yury Emelyanov, Alexey Morozov, Daniil Razdyakonov, and Mikhail Galkin. The simple web-based tool for visualization and sharing of semantic data and ontologies. In *International Semantic Web Conference (Posters & Demos)*, 10 2015.
- [61] Eero Hyvönen. *Publishing and using cultural heritage linked data on the semantic web*, volume 3. Morgan & Claypool Publishers, 2012.
- [62] Amit Singhal et al. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.
- [63] Kailash A Hambarde and Hugo Proenca. Information retrieval: recent advances and beyond. *IEEE Access*, 11:76581–76604, 2023.
- [64] Kumaripaba Athukorala, Dorota Głowacka, Giulio Jacucci, Antti Oulasvirta, and Jilles Vreeken. Is exploratory search different? a comparison of information search behavior for exploratory and lookup tasks. *Journal of the Association for Information Science and Technology*, 67(11):2635–2651, 2016.

- [65] Nicholas J Belkin. Anomalous states of knowledge as a basis for information retrieval. *Canadian journal of information science*, 5(1):133–143, 1980.
- [66] Gary Marchionini. Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4):41–46, April 2006. ISSN 0001-0782, 1557-7317. doi: 10.1145/1121949.1121979. URL <https://dl.acm.org/doi/10.1145/1121949.1121979>.
- [67] Ryen W White and Resa A Roth. *Exploratory search: Beyond the query-response paradigm*. Number 3. Morgan & Claypool Publishers, 2009.
- [68] Peter Pirolli and Stuart Card. Information foraging in information access environments. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 51–58, 1995.
- [69] John Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285, 1988.
- [70] Florian Windhager, Paolo Federico, Günther Schreder, Katrin Glinka, Marian Dörk, Silvia Miksch, and Eva Mayr. Visualization of cultural heritage collection data: State of the art and future challenges. *IEEE transactions on visualization and computer graphics*, 25(6):2311–2330, 2018.
- [71] Elaine G Toms et al. Serendipitous information retrieval. In *DELOS*, 2000.
- [72] Marcia J Bates. The design of browsing and berrypicking techniques for the online search interface. *Online review*, 13(5):407–424, 1989.
- [73] Marwan Al-Tawil, Vania Dimitrova, and Dhavalkumar Thakker. Using knowledge anchors to facilitate user exploration of data graphs. *Semantic Web*, 11(2):205–234, 2020.
- [74] Marwan Al-Tawil, Vania Dimitrova, Dhavalkumar Thakker, and Bilal Abu-Salih. Emerging Exploration Strategies of Knowledge Graphs. *IEEE Access*, 11:94713–94731, 2023. ISSN 2169-3536. doi: 10.1109/ACCESS.2023.3308514. URL <https://ieeexplore.ieee.org/document/10229152/>.
- [75] Yanji Chen, Mieczyslaw M. Kokar, and Jakub J. Moskal. SPARQL Query Generator (SQG). *Journal on Data Semantics*, 10(3):291–307, December 2021. ISSN 1861-2040. doi: 10.1007/s13740-021-00133-y. URL <https://doi.org/10.1007/s13740-021-00133-y>.
- [76] Hernán Vargas, Carlos Buil-Aranda, Aidan Hogan, and Claudia López. Rdf explorer: A visual sparql query builder. In *International Semantic Web Conference*, pages 647–663. Springer, 2019.
- [77] Axel-Cyrille Ngonga Ngomo, Lorenz Bühmann, Christina Unger, Jens Lehmann, and Daniel Gerber. Sorry, i don’t speak sparql: translating sparql queries into natural language. In *Proceedings of the 22nd International Conference on World Wide Web*,

- WWW '13, page 977–988, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450320351. doi: 10.1145/2488388.2488473. URL <https://doi.org/10.1145/2488388.2488473>.
- [78] Sébastien Ferré. Squall: a controlled natural language as expressive as sparql 1.1. In *International conference on application of natural language to information systems*, pages 114–125. Springer, 2013.
- [79] David Bawden and Lyn Robinson. The dark side of information: overload, anxiety and other paradoxes and pathologies. *Journal of Information Science*, 35(2): 180–191, 2009. doi: 10.1177/0165551508095781. URL <https://doi.org/10.1177/0165551508095781>.
- [80] Boris Katz. Annotating the world wide web using natural language. In *RIAO*, volume 1997, pages 136–159, 1997.
- [81] Vanessa Lopez, Miriam Fernández, Enrico Motta, and Nico Stieler. Poweraqua: Supporting users in querying and exploring the semantic web. *Semantic web*, 3(3): 249–265, 2012.
- [82] Vanessa Lopez, Christina Unger, Philipp Cimiano, and Enrico Motta. Evaluating question answering over linked data. *Journal of Web Semantics*, 21:3–13, 2013. ISSN 1570-8268. doi: <https://doi.org/10.1016/j.websem.2013.05.006>. URL <https://www.sciencedirect.com/science/article/pii/S157082681300022X>.
- [83] Andrei Broder. A taxonomy of web search. In *ACM Sigir forum*, volume 36, pages 3–10. ACM New York, NY, USA, 2002.
- [84] David P. Ausubel. A subsumption theory of meaningful verbal learning and retention. *The Journal of General Psychology*, 66(2):213–224, 1962. doi: 10.1080/00221309.1962.9711837. URL <https://doi.org/10.1080/00221309.1962.9711837>. PMID: 13863333.
- [85] Edwin L Hutchins, James D Hollan, and Donald A Norman. Direct manipulation interfaces. In *User centered system design*, pages 87–124. CRC Press, 1986.
- [86] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343, 1996. doi: 10.1109/VL.1996.545307.
- [87] Konstantinos Tsiakas and Dave Murray-Rust. Unpacking human-ai interactions: From interaction primitives to a design space. *ACM Transactions on Interactive Intelligent Systems*, 14(3):1–51, 2024.
- [88] Rafael Rocha and Gercina Ângela de Lima. Deigmata: A user-centered visualization tool for knowledge graph exploration. *Knowl. Organ.*, 52(8), 2025. doi: 10.31083/KO46137.

- [89] T. Ehrhart, P. Lisena, and R. Troncy. Kg explorer: A customisable exploration tool for knowledge graphs. In *VOILA 2021, International Workshop on the Visualization and Interaction for Ontologies and Linked Data*, 2021.
- [90] Weihao Jiang, Li Yan, Yaofeng Tu, Xiangsheng Zhou, and Zongmin Ma. Pg-explorer: Resource description framework data exploration with property graphs. *Expert Systems with Applications*, 198:116789, 2022.
- [91] Matteo Lissandrini, Davide Mottin, Katja Hose, and {Torben Bach} Pedersen. Knowledge graph exploration systems: are we lost? In *Proceedings of the 12th Conference on Innovative Data Systems Research*, January 2022. Annual Conference on Innovative Data Systems Research, CIDR ; Conference date: 09-01-2022 Through 12-01-2022.
- [92] Francisca Suárez and Aidan Hogan. Templet: a collaborative system for knowledge graph question answering over wikidata. In *Companion Proceedings of the ACM Web Conference 2023*, pages 152–155, 2023.
- [93] Alberto Moya Loustaunau and Aidan Hogan. Qawiki: A knowledge graph question answering & sparql query generation dataset for wikidata. 2025.
- [94] Mitchell Whitelaw. Generous interfaces for digital cultural collections. *Digital Humanities Quarterly*, 9(1):1–16, 2015. ISSN 1938-4122.
- [95] Paula Goodale, Paul David Clough, Samuel Fernando, Nigel Ford, and Mark Stevenson. Cognitive styles within an exploratory search system for digital libraries. *Journal of Documentation*, 70(6):970–996, 2014.
- [96] Drew Hemment, Cory Kommers, Ruth Ahnert, Maria Antoniak, Glauco Arbix, Vaishak Belle, Steve Benford, Alexandra Brintrup, Nick Bryan-Kinns, Mercedes Bunz, Baptiste Caramiaux, Sougwen Chung, Martin Disley, Yali Du, {Edgar A.} Duéñez-Guzman, Evelyn Gius, {Francisco Gómez} Medina, Lauren Goodlad, Naama Ilany-Tzur, Leif Isaksen, Marina Jirotko, Helen Kennedy, David Leslie, Dalaki Livingstone, Hoyt Long, Meredith Martin, Johanna Nalau, Chris Nathan, Ashley Noel-Hirst, Kirsten Ostherr, Andrew Prah, Omer Rana, Matt Ratto, Tobias Revell, Jenny Rhee, Isaac Rutenberg, Brent Seales, Stephanie Sherman, {Richard Jean} So, Adam Sobey, Jack Stilgoe, Marion Thain, Elaine Ubalijoro, Ted Underwood, Aditya Vashishta, Matjaz Vidmar, Matthew Wilkens, Youyou Wu, Sizhe Yuen, and Martin Zeilinger. *Doing AI differently: rethinking the foundations of AI via the humanities*. The Alan Turing Institute, July 2025. doi: 10.5281/zenodo.1642129. Licence: CC BY 4.0.
- [97] Reham Alharbi, Valentina Tamma, Floriana Grasso, and Terry R. Payne. The role of generative ai in competency question retrofitting. In Albert Meroño Peñuela, Oscar Corcho, Paul Groth, Elena Simperl, Valentina Tamma, Andrea Giovanni Nuzolese, Maria Poveda-Villalón, Marta Sabou, Valentina Presutti, Irene Celino, Artem

- Revenko, Joe Raad, Bruno Sartini, and Pasquale Lisena, editors, *The Semantic Web: ESWC 2024 Satellite Events*, pages 3–13, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-78952-6.
- [98] Dario Di Nuzzo, Edlira Vakaj, Hadeel Saadany, Eglantina Grishti, and Nandana Mihindukulasooriya. Automated generation of competency questions using large language models and knowledge graphs. In *SEMANTiCS Conference*, 2024.
 - [99] Youssra Rebboud, Lionel Tailhardat, Pasquale Lisena, and Raphael Troncy. Can llms generate competency questions? In *The Semantic Web: ESWC 2024 Satellite Events: Hersonissos, Crete, Greece, May 26–30, 2024, Proceedings, Part I*, page 71–80, Berlin, Heidelberg, 2025. Springer-Verlag. ISBN 978-3-031-78951-9. doi: 10.1007/978-3-031-78952-6_7. URL https://doi.org/10.1007/978-3-031-78952-6_7.
 - [100] Yousouf Taghzouti, Franck Michel, Tao Jiang, Louis-Félix Nothias, and Fabien Gandon. *Q²Forge: Minting Competency Questions and SPARQL Queries for Question-Answering Over Knowledge Graphs*. May 2025. doi: 10.48550/arXiv.2505.13572.
 - [101] Reham Alharbi, Jacopo de Berardinis, Floriana Grasso, T Payne, and Valentina Tamma. Characteristics and desiderata for competency question benchmarks. In *The Semantic Web-ISWC*, 2024.
 - [102] Vasile Ionut Remus Iga and Gheorghe Cosmin Silaghi. Assessing llms suitability for knowledge graph completion, 2024. URL <https://arxiv.org/abs/2405.17249>.
 - [103] Xiangqing Shen, Fanfan Wang, and Rui Xia. Reason-align-respond: Aligning llm reasoning with knowledge graphs for kgqa. *arXiv preprint arXiv:2505.20971*, 2025.
 - [104] Hoang Anh Dang, Vu Tran, and Le-Minh Nguyen. Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior. *Frontiers in Artificial Intelligence*, 8:1622292, 2025.
 - [105] Larissa Pusch, Alexandre Courtiol, and Tim Conrad. A human-in-the-loop, llm-centered architecture for knowledge-graph question answering. *arXiv preprint arXiv:2602.05512*, 2026.
 - [106] Ben Shneiderman. Human-centered artificial intelligence: Reliable, safe trustworthy. *International Journal of Human-Computer Interaction*, 36:1–10, 03 2020. doi: 10.1080/10447318.2020.1741118.
 - [107] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. volume 2015, 03 2015. doi: 10.1145/2678025.2701399.
 - [108] Streamlit Inc. Streamlit: The fastest way to build and share data apps. Web framework, 2023. URL <https://streamlit.io>. [Online; Accessed January 2026].

- [109] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.
- [110] Timothy Lebo, Satya Sahoo, Deborah Mcguinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. *PROV-O: The PROV Ontology*. 04 2013.
- [111] Andy Crabtree. H is for human and how (not) to evaluate qualitative research in hci. *Human-Computer Interaction*, 0(0):1–24, 2025. doi: 10.1080/07370024.2025.2475743. URL <https://doi.org/10.1080/07370024.2025.2475743>.
- [112] John W Creswell. *A concise introduction to mixed methods research*. SAGE publications, 2021.
- [113] Sandra G Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908. Sage publications Sage CA: Los Angeles, CA, 2006.
- [114] Luca Longo, Fabio Rusconi, Lucia Noce, and Stephen Barrett. The importance of human mental workload in web design. 2012.
- [115] Melanie J McGrath, Oliver Lack, James Tisch, and Andreas Duenser. Measuring trust in artificial intelligence: validation of an established scale and its short form. *Frontiers in Artificial Intelligence*, 8:1582880, 2025.
- [116] Ted Boren and Judith Ramey. Thinking aloud: Reconciling theory and practice. *IEEE transactions on professional communication*, 43(3):261–278, 2000.
- [117] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*, 2023.
- [118] Terry Koo and Mae Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15, 03 2016. doi: 10.1016/j.jcm.2016.02.012.
- [119] Nils Reimers and Iryna Gurevych. all-minilm-l6-v2. <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>, 2019. [Online; Accessed October 2025].
- [120] Nils Reimers and Iryna Gurevych. all-mpnet-base-v2. <https://huggingface.co/sentence-transformers/all-mpnet-base-v2>, 2021. Online; Accessed June 2025.
- [121] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta (roberta large), 2019. URL <https://huggingface.co/roberta-large>.
- [122] DeepSeek-V3.1. <https://api-docs.deepseek.com/news/news250821>, 2025. [Online; Accessed October 2025].

- [123] Qwen2.5 72B Instruct. <https://openrouter.ai/qwen/qwen-2.5-72b-instruct>, 2024. [Online; Accessed October 2025].
- [124] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611, 1965.
- [125] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. routledge, 2013.
- [126] William Jay Conover. *Practical nonparametric statistics*. john wiley & sons, 1999.
- [127] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.
- [128] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.
- [129] Claire McNamara. Tús maith - automatic generation of template competency questions, Jun 2025. URL osf.io/2xyn4.
- [130] Rensis Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 140:1–55, 1932.
- [131] Loris Grether and Hans Friedrich Witschel. Studying interaction patterns for knowledge graph exploration. In *KDIR*, pages 257–264, 2022.
- [132] Melike Sah and Vincent Wade. Personalized Concept-Based Search and Exploration on the Web of Data Using Results Categorization. In Philipp Cimiano, Oscar Corcho, Valentina Presutti, Laura Hollink, and Sebastian Rudolph, editors, *The Semantic Web: Semantics and Big Data*, Lecture Notes in Computer Science, pages 532–547, Berlin, Heidelberg, 2013. Springer. ISBN 978-3-642-38288-8. doi: 10.1007/978-3-642-38288-8_36.
- [133] Mohsen Tavakol and Reg Dennick. Making sense of cronbach’s alpha. *International journal of medical education*, 2:53, 2011.
- [134] Joseph F. Hair, G. Tomas M. Hult, Christian M. Ringle, Marko Sarstedt, Nicholas P. Danks, and Soumya Ray. Evaluation of Reflective Measurement Models. In *Partial Least Squares Structural Equation Modeling (PLS-SEM) Using R: A Workbook*, pages 75–90. Springer International Publishing, Cham, 2021. ISBN 978-3-030-80519-7. doi: 10.1007/978-3-030-80519-7_4. URL https://doi.org/10.1007/978-3-030-80519-7_4.
- [135] Klaas Sijtsma. On the use, the misuse, and the very limited usefulness of cronbach’s alpha. *Psychometrika*, 74(1):107–120, 2009.
- [136] Tenko Raykov. Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, 21(2):173–184, 1997.

- [137] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, 39(2):175–191, 2007.
- [138] Sang Gyu Kwak and Jong Hae Kim. Central limit theorem: the cornerstone of modern statistics. *Korean journal of anesthesiology*, 70(2):144, 2017.
- [139] Raquel V. Oliveira. *Type I and Type II Errors*, chapter 70, pages 348–353. John Wiley & Sons, Ltd, 2021. ISBN 9781119111931. doi: <https://doi.org/10.1002/9781119111931.ch70>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119111931.ch70>.
- [140] Mary-Jane Antia and C Maria Keet. Automating the generation of competency questions for ontologies with agocqs. In *Iberoamerican Knowledge Graphs and Semantic Web Conference*, pages 213–227. Springer, 2023.
- [141] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
- [142] Jennifer Edmond. *Digital Technology and the Practices of Humanities Research*. Open Book Publishers, 2020. ISBN 978-1-78374-839-6. doi: 10.11647/OBP.0192. URL <https://www.openbookpublishers.com/books/10.11647/obp.0192>.
- [143] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017. URL <https://arxiv.org/abs/1702.08608>.
- [144] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning, 2019. URL <https://arxiv.org/abs/1806.00069>.
- [145] Shreya Shankar, J. D. Zamfirescu-Pereira, Björn Hartmann, Aditya G. Parameswaran, and Ian Arawjo. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences, 2024. URL <https://arxiv.org/abs/2404.12272>.
- [146] U. Neisser. *Cognition and Reality: Principles and Implications of Cognitive Psychology*. Books in psychology. W. H. Freeman, 1976. ISBN 978-0-7167-0477-5. URL <https://books.google.ie/books?id=28u7QgAACAAJ>.
- [147] Sruti Subramanian, Katrien De Moor, Markus Fiedler, Kamil Koniuch, and Lucjan Janowski. Towards enhancing ecological validity in user studies: a systematic review of guidelines and implications for qoe research. *Quality and User Experience*, 8, 07 2023. doi: 10.1007/s41233-023-00059-2.
- [148] Yu Lu Liu, Su Lin Blodgett, Jackie Cheung, Q. Vera Liao, Alexandra Olteanu, and Ziang Xiao. ECBD: Evidence-centered benchmark design for NLP. In Lun-Wei

- Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16349–16365, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.861. URL <https://aclanthology.org/2024.acl-long.861/>.
- [149] John Foley, Brendan O'Connor, and James Allan. Improving entity ranking for keyword queries. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM '16*, page 2061–2064, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340731. doi: 10.1145/2983323.2983909. URL <https://doi.org/10.1145/2983323.2983909>.
 - [150] Alexander Ratner, Stephen H. Bach, Henry R. Ehrenberg, Jason Alan Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *CoRR*, abs/1711.10160, 2017. URL <http://arxiv.org/abs/1711.10160>.
 - [151] Holger Knublauch and Dimitris (eds.) Kontokostas. Shapes Constraint Language (SHACL). <https://www.w3.org/TR/shacl/>, 2017. [Online; Accessed October 2025].
 - [152] JSON Schema specification (2020-12). <https://json-schema.org/specification/>, 2020. [Online; Accessed October 2025].
 - [153] Google Analytics. <https://developers.google.com/analytics>, 2025. [Online; Accessed October 2025].
 - [154] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.
 - [155] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.
 - [156] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2, IJCAI'95*, page 1137–1143, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1558603638.
 - [157] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing Management*, 45:427–437, 07 2009. doi: 10.1016/j.ipm.2009.03.002.
 - [158] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine learning in python. *CoRR*, abs/1201.0490, 2012. URL <http://arxiv.org/abs/1201.0490>.

- [159] Zachary C. Lipton, Charles Elkan, and Balakrishnan Naryanaswamy. Optimal Thresholding of Classifiers to Maximize F1 Measure. In Toon Calders, Floriana Esposito, Eyke Hüllermeier, and Rosa Meo, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 225–239, Berlin, Heidelberg, 2014. Springer. ISBN 978-3-662-44851-9. doi: 10.1007/978-3-662-44851-9_15.
- [160] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. Distant supervision for relation extraction without labeled data. In Keh-Yih Su, Jian Su, Janyce Wiebe, and Haizhou Li, editors, *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore, August 2009. Association for Computational Linguistics. URL <https://aclanthology.org/P09-1113/>.
- [161] William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. A comparison of string distance metrics for name-matching tasks. In *Proceedings of the 2003 International Conference on Information Integration on the Web*, IIWEB’03, page 73–78. AAAI Press, 2003.
- [162] Wei Shen, Yuhan Li, Yinan Liu, Jiawei Han, Jianyong Wang, and Xiaojie Yuan. Entity linking meets deep learning: Techniques and solutions. *CoRR*, abs/2109.12520, 2021. URL <https://arxiv.org/abs/2109.12520>.
- [163] Greg Guest, Arwen Bunce, and Laura Johnson. How many interviews are enough? an experiment with data saturation and variability. *Field methods*, 18(1):59–82, 2006.
- [164] Paul J Lavrakas. *Encyclopedia of survey research methods*. Sage publications, 2008.
- [165] Jim Rudd, Ken Stern, and Scott Isensee. Low vs. high-fidelity prototyping debate. *interactions*, 3(1):76–85, 1996.
- [166] Peter Pirolli and Stuart Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, volume 5, pages 2–4. McLean, VA, USA, 2005.
- [167] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S Yu, and Tianyi Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment*, 4(11):992–1003, 2011.
- [168] Yizhou Sun and Jiawei Han. Mining heterogeneous information networks: a structural analysis approach. *ACM SIGKDD explorations newsletter*, 14(2):20–28, 2013.
- [169] Shixuan Liu, Changjun Fan, Kewei Cheng, Yunfei Wang, Peng Cui, Yizhou Sun, and Zhong Liu. Inductive meta-path learning for schema-complex heterogeneous information networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

- [170] Paul Alexandroff. *Elementary Concepts of Topology*, volume 747. Courier Corporation, 1961.
- [171] Alberto Cattaneo, Stephen Bonner, Thomas Martynec, Edward Morrissey, Carlo Luschi, Ian P Barrett, and Daniel Justus. The role of graph topology in the performance of biomedical knowledge graph completion models. *Bioinformatics*, 41(10):btaf547, 2025.

Appendix A

Appendix

A.1 Recruitment Emails

A.1.1 User Evaluation 1: Recruitment Email

Hi

One of my PhD students (Claire Mc Namara) is undertaking her research on how to engage users through curated questions and keywords related to people and places represented in an early draft of the knowledge graph explorer of the Virtual Record Treasury of Ireland which is due for release later this year.

The experiment does not require any previous experience of VRTI nor any particular expertise in Irish history, and should take approximately 20 minutes to complete

All the information that you need to complete the task can be found here:

<https://forms.gle/4R9FbUBqHadxdamP8>

Your participation would be greatly appreciated, and we are hoping you can complete the experiment by **February 14th** if at all possible. Thanks Declan

A.1.2 LLM Experiment: 1 Recruitment Email

Good afternoon,

I hope your week is going well!

The ethics application for the evaluation I have mentioned during the VRTI Tech meetings has just been approved. The evaluation consists of you both independently analysing two sets of competency questions for the VRTI KG that have been generated by two different Large Language Models.

Here is the link to the Google form for the evaluation:

<https://forms.gle/m2w7cntaeujboTkf9>. It should take less than 30 minutes to complete and everything relating to the evaluation is explained in the form. I have also attached a participant information leaflet to this email.

I am hoping that you both are able to complete the evaluation form by Monday, 02/06/2025. My apologies for the short turnaround, I am writing a paper about this experiment that needs to be completed for a Conference workshop by next Thursday.

Thank you for your time and if you have any questions, please feel free to reach out,

Kind regards, Claire

A.1.3 User Evaluation 2: Recruitment Email

A.1.3.1 Initial Recruitment Email

Hello,

I hope you are all doing well!

As some of you know, I am a final year PhD student (planning to submit February 2026) who has been working in the knowledge graph space to investigate how we can make large, complicated knowledge graphs more accessible to non-technical and non-domain-familiar people. I am using the Virtual Record Treasury of Ireland (VRTI) knowledge graph as a case study for my research.

The approach I have proposed and am investigating is the concept of using featured questions and answers as exploratory starting points into an unfamiliar knowledge graph. The idea is that these featured questions and answers will provide a jumping off point to begin exploring the knowledge graph, give users a sense of the kind of information they can find within the graph, and bring them to areas of the knowledge graph that they may not have otherwise thought to visit. As part of my research, I am developing an end-to-end pipeline to help domain experts (historians) in coming up with these featured questions, ensuring the questions are answerable by the graph, cover its scope, and that the entities the answers refer to are rich and not exploratory dead-ends. This pipeline and its tools are subject to a separate evaluation with historians to be undertaken later in the semester.

What I need your help with at this point however is to evaluate the concept of featured questions with people who are unfamiliar with the VRTI knowledge graph. This is to determine to what extent using these featured questions is useful to those people as exploratory starting points into a knowledge graph. Over the summer, I integrated the featured questions into the VRTI Explorer webpage, designed the user evaluation, put a questionnaire in place, and secured ethics approval to perform an evaluation of featured questions with students.

What I need from you at this stage therefore is the following: 1. While nearly complete, the pipeline to help historians generate these featured questions and answers is not yet ready to be used. I therefore will need 3 - 5 featured questions answers related to people and 3 - 5 featured questions and answers related to places that can be used as part of the evaluation. Such featured questions need to be related to (but not exclusively) content found within the knowledge graph as it currently exists and, for the answers to include entities that are as rich/interesting as possible. An example of a featured question could be the following: "What century was James Joyce active in?"

2. Identification of semester one history student cohorts (Peter mentioned MSc level students) to do the evaluation with and the method you feel would be the most appropriate for the delivery of the evaluation (by email, asking the students to do the evaluation in class, etc.). This is with the intention of garnering the maximum number of responses as possible (ideally 20+). Note: the students would not need briefing or exposure with respect to the VRTI knowledge graph, as the intention would be to get their blind responses to

the featured question/answer approach as a means to initially explore the VRTI knowledge graph.

I realise that this is a lot of information and I am sure that you will likely have many questions!

I have discussed the above with Peter already last week and he is positive that the above should in theory be doable.

Is there a particular day/time next week that would work for you all via zoom to have a meeting where I can demo the featured questions as part of the KG Explorer and answer any questions you may have about the above requests? I can do any time Monday except for between 12:30-1pm and any time Wednesday - Friday.

Thank you, I really appreciate your time,

Kind regards, Claire

A.1.3.2 Recruitment Email to History Professors

Hello,

Firstly, I would like to thank you all very much for your help in getting my user evaluation off the ground in the best shape possible!

Secondly, here is the link to the Google form that the students will use to take the evaluation: <https://forms.gle/7FzC1jda1tK3tg1M9>, and here is the link to the version of the VRTI Explorer that has featured questions implemented within it:

<https://mcnamacl.pythonanywhere.com/>

If you have a chance to look and have any feedback, please send your comments by Wednesday afternoon, as the evaluation is scheduled for Thursday.

Thank you again and I hope you all are having a lovely weekend,

Kind regards, Claire

A.1.4 User Evaluation 3: Recruitment Email

A.1.4.1 User Evaluation 3: VRTI Recruitment Email

Hello,

I hope this email finds you well!

I am reaching out to ask if you would participate in a 30-minute evaluation of a question-answer generation workflow I have designed as part of my PhD. You have been identified through your work with the VRTI project as someone who is an expert in history with relevant familiarity with the VRTI Knowledge Graph (<https://kg.virtualeconomy.ie/>).

My research to date has centred around investigating whether providing natural-language questions and answers is a useful way of guiding new users into the exploration of a knowledge graph (KG), and developing a workflow that enables domain experts (such as historians) to create such questions and answers.

If you are unfamiliar with my work to date, please watch the attached video demo and look at the attached screenshots so that you can see how such natural-language questions and answers might appear in the VRTI Knowledge Graph Explorer interface (so-called “Fea-

tured Questions”). You can also interact with the featured questions directly by following this link: <https://mcnamacl.pythonanywhere.com/>

For this evaluation, however, you will be asked to use a prototype interface that is intended to help historians create question-answer pairs like those you saw above (featured questions), based on information that exists in the VRTI KG. Ultimately, we are trying to determine if the provided workflow is one that you would find useful if tasked with creating featured questions and answers.

You will be asked to perform a series of tasks using the interface and to verbalise your thoughts as you do so. This will be audio-recorded with your consent. Following this, you will be asked to fill out a Google form that will ask you questions about your experience using the workflow and ask you to evaluate the questions and answers created through the interface.

The evaluation will take around 30 minutes and will take place online via Microsoft Teams. You can schedule a time by following this link: <https://calendly.com/mcnamacl-tcd/tus-maith-evaluation>. Once we have scheduled a time, I will send you an email with the Microsoft Teams meeting link and a link to the Google questionnaire you will be asked to fill out following the evaluation.

I have attached a participant information leaflet along with a consent form - both of which provide further information about the evaluation. You do not need to sign the attached consent form, as consent will be obtained digitally on the day of the evaluation. The form has been provided solely to give you an indication of what will be asked of you.

Thank you very much for your time and help - I know this is a busy time of year! Your insights are invaluable to my research and will contribute to the wider landscape of knowledge graph exploration research.

If you have any questions at all, please do not hesitate to contact me at my email address: mcnamacl@tcd.ie

Kind regards, Claire

A.1.4.2 User Evaluation 3: VOICES Recruitment Email

Hello,

I hope this email finds you well!

I am reaching out to ask if you would participate in a 30-minute evaluation of a question-answer generation workflow I have designed as part of my PhD. You have been identified through your work with the VOICES project as someone who is an expert in history with relevant familiarity with the VRTI Knowledge Graph (<https://kg.virtualtreasury.ie/>).

My research to date has centred around investigating whether providing natural-language questions and answers is a useful way of guiding new users into the exploration of a knowledge graph (KG), and developing a workflow that enables domain experts (such as historians) to create such questions and answers.

If you are unfamiliar with my work to date, please watch the attached video demo and look at the attached screenshots so that you can see how such natural-language questions and answers might appear in the VRTI Knowledge Graph Explorer interface (so-called “Fea-

tured Questions”). You can also interact with the featured questions directly by following this link: <https://mcnamacl.pythonanywhere.com/>

For this evaluation, however, you will be asked to use a prototype interface that is intended to help historians create question-answer pairs like those you saw above (featured questions), based on information that exists in the VRTI KG. Ultimately, we are trying to determine if the provided workflow is one that you would find useful if tasked with creating featured questions and answers.

You will be asked to perform a series of tasks using the interface and to verbalise your thoughts as you do so. This will be audio-recorded with your consent. Following this, you will be asked to fill out a Google form that will ask you questions about your experience using the workflow and ask you to evaluate the questions and answers created through the interface.

The evaluation will take around 30 minutes and will take place online via Microsoft Teams. You can schedule a time by following this link: <https://calendly.com/mcnamacl-tcd/tus-maith-evaluation>. Once we have scheduled a time, I will send you an email with the Microsoft Teams meeting link and a link to the Google questionnaire you will be asked to fill out following the evaluation.

I have attached a participant information leaflet along with a consent form - both of which provide further information about the evaluation. You do not need to sign the attached consent form, as consent will be obtained digitally on the day of the evaluation. The form has been provided solely to give you an indication of what will be asked of you.

Thank you very much for your time and help - I know this is a busy time of year! Your insights are invaluable to my research and will contribute to the wider landscape of knowledge graph exploration research.

If you have any questions at all, please do not hesitate to contact me at my email address: mcnamacl@tcd.ie

Kind regards, Claire

A.2 Questionnaires

A.2.1 LLM Experiment 1: Questionnaire

Evaluation Criteria (for R/C/A ratings) 1 - 5 Likert Scale (1=Not at all, 5=Very)

Relevance (R): In your opinion, how well does the question align with the purpose of competency questions?

Clarity (C): Is the question phrased in a clear, unambiguous, and grammatically correct way that reflects its intended meaning?

Answerability (C): In your opinion, can the knowledge graph reasonably provide an answer to the question?

Set 1: Questions 1-20

No.	Competency Question	Criteria (R/C/A)
1	What is the preferred identifier of a person?	R/C/A
2	What type of human-made thing is this?	R/C/A
3	What place is this part of?	R/C/A
4	What is the area of interest of a place?	R/C/A
5	What is the time-span of an event?	R/C/A
6	When did a person die?	R/C/A
7	When was a person born?	R/C/A
8	What are the dimensions of a physical thing?	R/C/A
9	What is the hectare size of a place?	R/C/A
10	What group is this person a member of?	R/C/A
11	What attribute was assigned to a person?	R/C/A
12	What is an alternative form of an appellation?	R/C/A
13	What place is a birth associated with?	R/C/A
14	What place is a death associated with?	R/C/A
15	What title does a person have?	R/C/A
16	What is depicted by a physical thing?	R/C/A
17	What document lists this?	R/C/A
18	What are the components of something?	R/C/A
19	What identifiers exist for a place?	R/C/A
20	Whom does a person report to?	R/C/A

Note: R = Relevance, C = Clarity, A = Answerability

Domain Expert Questions (Reference for Both Sets)

Here is a list of 10 domain expert generated competency questions:

1. When was a person born?
2. Where was a person born?
3. When did a person die?
4. Where did a person die?
5. What century was a person active in?
6. Which people have this surname?
7. What type of place is this?
8. What period of time/era does this placename belong to?
9. Does this person have a Dictionary of Irish Biography entry?
10. What county is X parish/townland/barony located in?

Are there any questions generated by the LLM that are not mentioned in the domain experts list of questions that you feel could be added?

Open-ended Questions for Set 1

1. If yes, please provide the question/s and a reason as to why you feel that it/they could be added to the domain experts list.
2. If you have any other comments about set one that you would like to add, please include them here.

Set 2: Questions 21-40

No.	Competency Question	Criteria (R/C/A)
21	Who carried out the attribute assignment?	R/C/A
22	What type of attribute assignment was carried out?	R/C/A
23	Who was assigned in the attribute assignment?	R/C/A
24	What value was assigned in the attribute assignment?	R/C/A
25	Which type of attribute was assigned to a person?	R/C/A
26	Who is identified by this identifier?	R/C/A
27	What is the gender of a person?	R/C/A
28	When was a person born?	R/C/A
29	When did a person die?	R/C/A
30	What role did a person have?	R/C/A
31	What is the geometric identifier of a place?	R/C/A
32	Who does the counselor work with?	R/C/A
33	What is the relationship between a group and a person?	R/C/A
34	Who reports to whom in the group?	R/C/A
35	Which symbolic object represents a group?	R/C/A
36	What elements compose this object?	R/C/A
37	What place is part of a larger area of interest?	R/C/A
38	What is the link between these subjects?	R/C/A
39	What is the main subject of the area of interest?	R/C/A
40	What is the century associated with an event?	R/C/A

Note: R = Relevance, C = Clarity, A = Answerability

Domain Expert Questions (Reference for Both Sets)

Here is a list of 10 domain expert generated competency questions:

1. When was a person born?
2. Where was a person born?
3. When did a person die?
4. Where did a person die?
5. What century was a person active in?
6. Which people have this surname?
7. What type of place is this?

8. What period of time/era does this placename belong to?
9. Does this person have a Dictionary of Irish Biography entry?
10. What county is X parish/townland/barony located in?

Are there any questions generated by the LLM that are not mentioned in the domain experts list of questions that you feel could be added?

Open-ended Questions for Set 2

1. If yes, please provide the question/s and a reason as to why you feel that it/they could be added to the domain experts list.
2. If you have any other comments about set two that you would like to add, please include them here.

Final Comparison Questions

1. Which set of questions did you prefer?
2. Please give a reason/s as to why you picked your preferred set.
3. If you have any other comments that you would like to add, please include them here.

A.3 User Evaluation Scripts

A.3.1 User Evaluation 3: Script

You have been asked to participate in this evaluation in your capacity as a domain expert historian with relevant familiarity of the VRTI knowledge graph. My PhD work to date has centred around investigating how we can provide new users with meaningful entry points into beginning the exploration of an unfamiliar knowledge graph. The solution I am proposing is a set of questions and answers that expose the scope of the knowledge graph and whose answers bring users to information-rich areas of it. I am firstly going to give you a very short demo of what these questions and answers, or “featured questions”, can look like in practice. The featured questions have been integrated into a prototype VRTI Explorer and the questions and answers you will see demonstrate the 3 kinds of featured questions and answers: Interface-dependent. Complex. Simple.

We are going to be focusing on the simple questions today (one-hop questions). I am then going to give you control of my screen in order to utilise a prototype interface. This interface is an early-stage workflow that makes use of AI and is designed to support domain experts in generating simple featured questions and answers that can act as entry points into the VRTI knowledge graph. The aim is to help domain experts create short, meaningful questions whose answers lead users into information-rich areas of the KG. Today we are evaluating the workflow itself, the three-stage process, rather than the visual interface or the specific outputs you will see. I am interested in your expert perspective on how viable this workflow feels, what works, what doesn’t, and how you reason through each step.

As mentioned, this interface is there only to give you an idea of what a workflow I have been developing might look like in practice. I will ask you to undertake a task using this interface and record the audio as you do so, in other words, I would ask you to speak aloud as you are working through the task mentioning anything that comes to mind - surprising, confusing, good, bad, anything! Finally, you will fill out a Google form questionnaire based on your experience. Again, I want to emphasise, you are not being asked to evaluate the user interface itself, but the process you are following.

<Demo of featured questions>

Task: Imagine it was the case you were tasked with coming up with a set of simple featured questions and answers and you were able to use the following work flow. Again, to be clear, we are not evaluating the interface itself that you are using to create these questions and answers (that can be changed/updated) but we are evaluating the provided workflow. In other words, the 3 stage approach shown to you comprising of template questions that are there to expose the scope of the knowledge graph, followed by a way to provide potential entities of interest to you whose information will fill in these template questions, followed by the reviewal of filled in questions and answers based on the templates and entities you have chosen. The prototype interface is merely there to demonstrate it in action and has not been heavily engineered.

With your consent, I will now turn on audio recording via my iPhone.

Please say everything you are thinking as you work. Choose 4 template questions that look suitable for the task at hand (which as a reminder, are simple questions that could act as entry points into the VRTI KG via the VRTI Explorer). Choose 2 place related questions and 2 person related questions.

The following search terms are the top 5 terms public users have been searching over the VRTI knowledge graph using the VRTI Explorer. Choose a search term to see the entities that match the search term.

The entities you see here are related to search term you chose previously. The list is ordered by an automatic relevance score that uses lexical matching and internal measures of how much information the KG holds for each entity Choose at least two person entities and two place entities whose information will be used to fill in the template questions you chose.

If the entity contains information that can be filled in to the template questions you chose, you will see the filled in question and answer.

Review the filled in questions and answers. Take as long as you need to inspect the generated questions and answers. Once you are satisfied that you have reviewed them to your liking, please press the button at the bottom that says “finished reviewing”.

After completing the task, I will ask you to reflect not on each individual question you saw, but on the capability of the workflow itself. I’ll ask for your overall judgment on whether a process like this could, in principle, help produce the kinds of entry-point questions we aim for.

A.4 Ethics Applications

A.4.1 User Evaluation 1: Ethics Application

Applicant & Collaborators

Applicant Details

- Applicant Name : Claire Mcnamara
- Is applicant applying as a member of staff or a student? : Student
- School / Department : School of Computer Science & Statistics - Computer Science
- Role on the Project : Principal Investigator
- Primary Employer (if not TCD) :
- Other affiliations (if applicable) :
- Course (for student applicant only) : PhD
- Part time / full time : Full-Time
- I have read and understood Trinity policy on Good Research Practice : Yes
- I have read and completed the Integrity module : Yes
- I have read and completed the Data Protection Training module within the last 24 months : No

Trinity Collaborators

User	Role
Declan O'Sullivan	Primary Supervisor
Lucy Hederman	Trinity Co supervisor

Other Collaborators

No data available in this section.

Project Details

Main Project Details

- Title of research :Tús Maith
- Data Collection Start Date : 13/01/2025
- Data Collection End Date : 28/02/2025
- Project End Date : 12/03/2025 12:00:00 AM
- Does the research involve :Humans (or their data)
- Does this Animal research involve :
- Is this AREC research a first application or an amendment :
- Is the research WHOLLY (i) an analysis of statutes/legal provisions AND/OR (ii) legal judgments/cases which have been made public by a judicial process? :
- Trinity Researchers will have access ONLY to aggregate data or data that otherwise fall outside of the remit of the GDPR. :
- Could the research have detrimental legal, economic or social consequences for either the participant or their establishments : No
- Intentions of the research: Does the research : None of the above
- State research aim(s) and objective(s), research question or hypothesis : The objective of this study is to identify issues in the design and implementation of the Tús Maith (TM). TM comprises of a set of features added to the existing VRTI explorer. It is hoped that the study gathers evidence to determine if users find the starting points provided helpful to them in commencing their exploration of the VRTI knowledge graph. The hypothesis of the study is “TM provides a new user with a useful starting point to begin their exploration of the VRTI knowledge graph”.
- Lay Summary: including background/ rationale/ justification, research approach, study design (exclude detail of measurement instruments and intervention and analysis (if applicable) (Word limit: 250 words) : Background: The Virtual Treasury of Ireland (VRTI) knowledge graph contains information about Irish historical data which was reconstructed from the remnants of a fire in the Public Records Office of Ireland, destroyed during the civil war in 1922. Interacting with knowledge graphs requires a high level of relevant technical background knowledge. Thus, an application was created to allow domain experts (e.g historians) with limited relevant technical background to search and edit entities in the knowledge graph. Tús Maith (Irish for “Good Start”) is an addition to this application that aims to ease new users into exploring the VRTI knowledge graph. Rationale: It is hoped that Tús Maith will provide new users with a useful starting off point. Justification: The study conducted via Google forms will allow feedback from end users to be captured , analyzed and used to refine the design of Tús Maith. Research Approach: The participants will interact with the application in accordance with questions provided in the questionnaire. Study Design: The study is designed to gather feedback from end users how they find the addition of extra features on top of the existing VRTI exploration interface. Thus, providing information which can be used to refine and validate the design of the exploration starting points. Briefing information for participants is made available as web pages once the participant begins the experiment. The web pages will provide access to the web application and associated documentation. All feedback received will be fully anonymous and no identifiable information will be stored.
- Identify all countries where data is collected or processed : IE

- Does the project involve :Human participants and /or their data and no biological samples
 - Is the project funded : No
-

Details on Human Participants and their Data

- Does your research involve Mixed Methods : No
 - Does the project use data from :Primary sources only
 - Will you obtain Consent from participants (or in the case of children or adults at risk of vulnerability from a parent / legal guardian) for their participation and/ or the use of their data? : Yes
 - Provide further information :
 - Will payment be made to research participants? :No
 - Provide further information :
 - Is the research Health Research as defined in the Health Research Regulations? : No
 - Does the research require a Consent Declaration form as defined by the Health Research Regulations 2018 and amendments 2021? :
 - Which of the following best describes the research Secondary Data Source(s)? :
 - Could any of the research data directly identify any participant? : No
 - Could any of the research data indirectly identify any participant? :No
 - Do you process Personal Data for study administration purposes? e.g. contacting individuals : No
 - If yes please outline how you will store the information, keep it confidential, and destroy it in line with the data retention policy. :
 - Which of the following best describes the general characteristics of the target population? :Students of Trinity
 - Do any of the following further describe the characteristics of the target population? :
 - Please describe :
 - List the inclusion/ exclusion criteria for selection of research participants : Inclusion: Participants will be students with an interest in History. Exclusion: No participants under the age of 18 will be recruited.
-

Research Sources and Sites

Name	Req
Google forms	NOT-APPLICABLE

Outline of research methods

No data available in this section.

-
- Does your research use any of the following methods exclusively? : Anonymous Surveys (Note. Only select if there are no free text responses in the survey)
 - Does the research include an intervention? : No
 - Please select which of the following best describes the intervention :
 - Other :
 - Outline the intervention :
 - Outline the method of analysis (Word limit :100 words) : The method of analysis will involve conducting thematic analysis on the qualitative data received from the participants. The analysis will result in the detection of patterns which can guide the improvements of Tús Maith. Each discovered pattern will be reviewed and implemented if deemed necessary. In addition, the quantitative data will be compared with acceptable research thresholds to identify possible areas for improvement.
 - What is the approximate size of the target population? : 30
 - What is the proposed sample size -how many participants are involved in the study? : 15
 - What is the justification for the sample size? : This is the minimum number of qualified participants needed to understand if the added features are useful as part of Tús Maith.
-
-

Risk

Risk , benefit, harm to researcher

- What setting/s will be used for data collection : Online
- Other elaborate :

Type	Other Type	Impact	Probabilty	Mitigation
None				

Risk , benefit, harm to site environment or Society

Type	Impact	Probabilty	Mitigation
None			

Risk , benefit, harm to participants

- Do you or any of the research team have any dependant relationship to the participants? : : No

- Detail the role of the researcher/s, the relationship of the researcher/s to the participants normally, and how you will mitigate against coercion of participants to participate in the research or in influencing their responses :

Type	Other Type	Impact	Probability	Mitigation
None				

Participant Benefits

- Is it foreseeable that participants could potentially reveal information that you have a legal obligation to disclose (e.g. child protection policy, malpractice, etc.) :No

- What information may be disclosed, why, and to whom. :

- Outline any direct benefits to research participants of participation : To progress research related to a tool which can be used in their domain.

Conflict of Interest

- Are you aware of any Conflict of Interest from the PI or any collaborator, processor or other person involved in the conduct of the project, that could arise in the course of the project? : No

- Give details of the Conflict of Interest and what mitigation measures are in place :

Consent

Consent

- Do your participants require support to give Consent : No

- How will Assent be obtained and by whom :

- Provide Further Information :

- Do you require Assent or Proxy Consent from participants e.g. because of their vulnerability : No

- Provide Further Information :

- Are you required to have Garda Clearance : No

- What is the time interval between giving information and securing Consent : 7 or More Days

- Provide Further Information :

- Describe how you will inform participants about the use of their Personal Data : Participants will be made aware prior to the commencement of the experiment that their anonymised data may be used in publications related to the work as direct quotations. No identifiable information will be stored,

therefore, no identifying information will be contained in the publications. This information will be provided in the informed consent.

- Describe how participants can withdraw their Consent and/or their data : The study is fully anonymous, therefore, the confirmation page will state that no personal data has been recorded and that the study is fully anonymous.

Sampling & Recruitment

Sampling and Recruitment

- Outline the sampling method : The population for consideration will be the students with an interest in History. Then, voluntary response sampling will be used to recruit participants who would be willing to participate in the study.

- Describe the time commitment of the participant : The questionnaire should take no longer than twenty minutes to complete.

- Will the research require/use a gatekeeper :No

- Outline the position/role of the gatekeeper within the organisation :

- Detail the role of the gatekeeper in the research :

- Is there a dependant relationship between the gatekeeper and the participants :

- Outline how this is going to be managed to mitigate against the dependencies :

- Please give a detailed step by step description of how participants will be recruited and append the recruitment material (Word limit: 100 words). : The students who meet the inclusion criteria will be invited using the email shown in the uploaded recruitment documentation. Those who agree to participate will be sent follow up information, detailing how to complete the experiment.

A.4.2 User Evaluation 2: Ethics Application

Applicant & Collaborators

Applicant Details

- Applicant Name : Claire Mcnamara
- Is applicant applying as a member of staff or a student? : Student
- School / Department : School of Computer Science & Statistics - Computer Science
- Role on the Project : Principal Investigator
- Primary Employer (if not TCD) :
- Other affiliations (if applicable) :
- Course (for student applicant only) : PhD
- Part time / full time : Full-Time
- I have read and understood Trinity policy on Good Research Practice : Yes
- I have read and completed the Integrity module : Yes
- I have read and completed the Data Protection Training module within the last 24 months : No

Trinity Collaborators

User	Role
Lucy Hederman	Primary Supervisor
Declan O'Sullivan	Trinity Co supervisor

Other Collaborators

No data available in this section.

Project Details

Main Project Details

- Title of research :Tús Maith - User Evaluation 2
- Data Collection Start Date : 18/08/2025
- Data Collection End Date : 18/09/2025
- Project End Date : 31/10/2025 12:00:00 AM
- Does the research involve :Humans (or their data)
- Does this Animal research involve :
- Is this AREC research a first application or an amendment :
- Is the research WHOLLY (i) an analysis of statutes/legal provisions AND/OR (ii) legal judgments/cases which have been made public by a judicial process? :
- Trinity Researchers will have access ONLY to aggregate data or data that otherwise fall outside of the remit of the GDPR. :
- Could the research have detrimental legal, economic or social consequences for either the participant or their establishments : No
- Intentions of the research: Does the research : None of the above
- State research aim(s) and objective(s), research question or hypothesis : This research aims to evaluate whether curated natural language questions can support non-technical users in beginning meaningful exploration of the Virtual Record Treasury of Ireland Knowledge Graph (VRTI KG). The study investigates whether these structured entry points reduce initial uncertainty, foster confidence, and help users form a conceptual understanding of the available content. The research question is: To what extent do curated questions help users overcome the Initial Exploration Problem when interacting with complex historical knowledge graphs? The objective is to assess user perceptions of usefulness, clarity, and confidence through a lightweight, questionnaire-based usability evaluation.
- Lay Summary: including background/ rationale/ justification, research approach, study design (exclude detail of measurement instruments and intervention and analysis (if applicable) (Word limit: 250 words) : The VRTI KG is a large and complex KG that reconstructs Irish historical records lost in the destruction of the Public Record Office of Ireland in 1922. While it offers rich historical content, its complex structure can pose challenges for non-technical users who are unfamiliar with semantic technologies or large-scale digital archives. Previous research identified the “Initial Exploration Problem,” which refers to the difficulty users face when first trying to explore unfamiliar and complex knowledge graphs. This study evaluates whether a specific design feature, curated questions, can help address this problem by guiding users into meaningful exploration pathways. Curated questions are structured prompts shown at the beginning of a user’s interaction with the system (e.g. “Who were the people involved in X event?”). These questions are designed to help users understand what kinds of information the system contains and how they might begin exploring it. Participants will be asked to interact with the curated questions via the VRTI Explorer (an interface for the VRTI KG) in their own time via a public webpage. They will be instructed to select a person-based and a place-based curated question, and explore freely from there. After completing their session, they will be invited to fill out a short questionnaire about their experience. The aim of this research is to understand whether curated questions help non-technical users overcome the barriers of first-time use and increase their confidence when exploring the archive. Insights from this study will inform the future design of accessible interfaces for cultural heritage KGs.

- Identify all countries where data is collected or processed : IE
 - Does the project involve :Human participants and /or their data and no biological samples
 - Is the project funded : Yes
-

Details on Human Participants and their Data

- Does your research involve Mixed Methods : No
 - Does the project use data from :Primary sources only
 - Will you obtain Consent from participants (or in the case of children or adults at risk of vulnerability from a parent / legal guardian) for their participation and/ or the use of their data? : Yes
 - Provide further information :
 - Will payment be made to research participants? :No
 - Provide further information :
 - Is the research Health Research as defined in the Health Research Regulations? : No
 - Does the research require a Consent Declaration form as defined by the Health Research Regulations 2018 and amendments 2021? :
 - Which of the following best describes the research Secondary Data Source(s)? :
 - Could any of the research data directly identify any participant? : No
 - Could any of the research data indirectly identify any participant? :No
 - Do you process Personal Data for study administration purposes? e.g. contacting individuals : No
 - If yes please outline how you will store the information, keep it confidential, and destroy it in line with the data retention policy. :
 - Which of the following best describes the general characteristics of the target population? :Students of Trinity
 - Do any of the following further describe the characteristics of the target population? :
 - Please describe :
 - List the inclusion/ exclusion criteria for selection of research participants : TCD students with an interest in Irish History taking a relevant module run by a professor working with or adjacent with the VRTI Project.
-

Research Sources and Sites

Name	Req
Google forms	NOT-APPLICABLE

Outline of research methods

No data available in this section.

-
- Does your research use any of the following methods exclusively? : Information, documents or data which are in the public domain
 - Does the research include an intervention? : No
 - Please select which of the following best describes the intervention :
 - Other :
 - Outline the intervention :
 - Outline the method of analysis (Word limit :100 words) : Participant responses to the questionnaire will be analysed using both quantitative and qualitative methods. Likert-scale questions will assess perceived confidence, ease of use, and the perceived value of curated questions in supporting exploration. Descriptive statistics (means and standard deviations) will be used to identify overall trends in user perception. Open-text responses will undergo thematic analysis to identify recurring patterns related to user confidence, perceived usefulness, and the role of curated questions in shaping early interaction. This approach is designed to evaluate the extent to which curated questions help mitigate the Initial Exploration Problem.
 - What is the approximate size of the target population? : 60
 - What is the proposed sample size -how many participants are involved in the study? : 30
 - What is the justification for the sample size? : The number of eligible students in TCD that can be recruited through professors working with or adjacent with the VRTI Project.
-
-

Risk

Risk , benefit, harm to researcher

- What setting/s will be used for data collection : Place of convenience for participants
- Other elaborate :

Type	Other Type	Impact	Probability	Mitigation
None				

Risk , benefit, harm to site environment or Society

Type	Impact	Probability	Mitigation
None			

Risk , benefit, harm to participants

- Do you or any of the research team have any dependant relationship to the participants? : : No
- Detail the role of the researcher/s, the relationship of the researcher/s to the participants normally, and how you will mitigate against coercion of participants to participate in the research or in influencing their responses :

Type	Other Type	Impact	Probability	Mitigation
None				

Participant Benefits

- Is it foreseeable that participants could potentially reveal information that you have a legal obligation to disclose (e.g. child protection policy, malpractice, etc.) :No
- What information may be disclosed, why, and to whom. :
- Outline any direct benefits to research participants of participation : Participants may benefit by gaining early access to a new way of exploring Irish historical records and by contributing to research that aims to make the complex KG more accessible and easier to navigate for the public.

Conflict of Interest

- Are you aware of any Conflict of Interest from the PI or any collaborator, processor or other person involved in the conduct of the project, that could arise in the course of the project? : No
- Give details of the Conflict of Interest and what mitigation measures are in place :

Consent

Consent

- Do your participants require support to give Consent : No
- How will Assent be obtained and by whom :
- Provide Further Information :
- Do you require Assent or Proxy Consent from participants e.g. because of their vulnerability : No
- Provide Further Information :
- Are you required to have Garda Clearance : No
- What is the time interval between giving information and securing Consent : 7 or More Days

- Provide Further Information :

- Describe how you will inform participants about the use of their Personal Data : Participants will be made aware prior to the commencement of the experiment that their anonymised data may be used in publications related to the work as direct quotations. No identifiable information will be stored, therefore, no identifying information will be contained in the publications. This information will be provided in the informed consent.

- Describe how participants can withdraw their Consent and/or their data : Once participants have completed the questionnaire and have submitted their answers it will no longer be possible to identify or erase their data. This is due to the fact that the data is anonymous. The research team has no means of distinguishing their data from those of any of the other participants. There will be a final question just before submission that will ask if the participant is happy to submit their answers or not and an exit message reiterating their chosen option.

Funding

Funding

- Insert RPAMS number if applicable and available :

- Outline sources of funding, list names of all confirmed sources of funding or support (including in-kind benefit). For all, state if they are industry/commercial, state/public, philanthropic/charitable, other : The research conducted was funded by the Irish Research Council under project ID GOIPG/2021/1165 and supported by the Research Ireland ADAPT Centre for Digital Content Technology (grant number 13/RC/2106 P2).

- Please specify any funder specific requirements or obligations which should be brought to the attention of the ethics committee and or Trinity Innovation & Enterprise : There are no such requirements.

- Will the results of the project be used or disclosed for commercial purposes : No

- Please clarify which party shall have the commercialisation and/or intellectual property rights : The Irish Research Council and the ADAPT Research Centre.

Conflict of Interest

- Are you aware of any possible Conflict of Interest arising from the funding or commercialisation of this project : : No

- Give details of the Conflict of Interest and what mitigation measures are in place :

- It there likely to be possible Conflict of Interest between the funders of the project and the aims and results of the project : No

- Give details of the Conflict of Interest and what mitigation measures are in place :

Sampling & Recruitment

Sampling and Recruitment

- Outline the sampling method : The population for consideration will be TCD students with an interest in History taking a module taught by a professor working with or adjacent to the VRTI Project. Then, voluntary response sampling will be used to recruit participants who would be willing to participate in the study.
 - Describe the time commitment of the participant : The questionnaire should take no longer than 10 minutes to complete.
 - Will the research require/use a gatekeeper :No
 - Outline the position/role of the gatekeeper within the organisation :
 - Detail the role of the gatekeeper in the research :
 - Is there a dependant relationship between the gatekeeper and the participants :
 - Outline how this is going to be managed to mitigate against the dependencies :
 - Please give a detailed step by step description of how participants will be recruited and append the recruitment material (Word limit: 100 words). : The students who meet the inclusion criteria will be invited via an email sent by TCD professors working with or adjacent with the VRTI Project. The email will contain the link to the questionnaire that they may complete at a time that is convenient to them.
-

A.4.3 LLM Experiment 1: Ethics Application

Applicant & Collaborators

Applicant Details

- Applicant Name : Claire Mcnamara
- Is applicant applying as a member of staff or a student? : Student
- School / Department : School of Computer Science & Statistics - Computer Science
- Role on the Project : Principal Investigator
- Primary Employer (if not TCD) :
- Other affiliations (if applicable) :
- Course (for student applicant only) : PhD
- Part time / full time : Full-Time
- I have read and understood Trinity policy on Good Research Practice : Yes
- I have read and completed the Integrity module : Yes
- I have read and completed the Data Protection Training module within the last 24 months : No

Trinity Collaborators

User	Role
Lucy Hederman	Trinity Co supervisor
Declan O'Sullivan	Primary Supervisor

Other Collaborators

No data available in this section.

Project Details

Main Project Details

- Title of research :Tús Maith - Automatic Competency Question Generation
- Data Collection Start Date : 16/05/2025
- Data Collection End Date : 30/05/2025
- Project End Date : 20/06/2025 12:00:00 AM
- Does the research involve :Humans (or their data)
- Does this Animal research involve :
- Is this AREC research a first application or an amendment :
- Is the research WHOLLY (i) an analysis of statutes/legal provisions AND/OR (ii) legal judgments/cases which have been made public by a judicial process? :
- Trinity Researchers will have access ONLY to aggregate data or data that otherwise fall outside of the remit of the GDPR. :
- Could the research have detrimental legal, economic or social consequences for either the participant or their establishments : No
- Intentions of the research: Does the research : None of the above
- State research aim(s) and objective(s), research question or hypothesis : This study explores the extent to which an open-source Large Language Model (LLM) can generate competency questions (CQs) to support Tús Maith (TM), a multistep framework designed to help new users explore the VRTI knowledge graph. The research question is: To what extent can an LLM generate reasonable CQs using the VRTI KG for the purposes of TM? Here, a CQ is defined as a natural language question that expresses a query the knowledge graph should be able to answer, helping to convey its scope and potential to users unfamiliar with its structure.
- Lay Summary: including background/ rationale/ justification, research approach, study design (exclude detail of measurement instruments and intervention and analysis (if applicable) (Word limit: 250 words) : Background: The Virtual Treasury of Ireland (VRTI) knowledge graph is a large and diverse KG that contains information about Irish historical data which was reconstructed from the remnants of a fire in the Public Records Office of Ireland, destroyed during the civil war in 1922. Rationale: In order to assess how reasonable the competency questions generated by an LLM are, they need to be evaluated by a domain expert(s). This is to assess to what extent the prompt and LLM used can generate valid and reasonable competency questions for the VRTI KG. Research approach: participants will interact with a Google form in their own time assessing each question. Study design: This evaluation aims to evaluate the effectiveness of an LLM in generating competency questions for the VRTI KG. The findings will inform potential refinements to the prompt and add a validated set of competency questions for future research. Participants will receive a detailed briefing via email, the Google form, and a participant information form. Each participant will be provided a link to the Google form containing the generated questions along with a structured evaluation section to assess each question based on predefined criteria.
- Identify all countries where data is collected or processed : IE
- Does the project involve :Human participants and /or their data and no biological samples
- Is the project funded : Yes

Details on Human Participants and their Data

- Does your research involve Mixed Methods : No
- Does the project use data from :Primary sources only
- Will you obtain Consent from participants (or in the case of children or adults at risk of vulnerability from a parent / legal guardian) for their participation and/ or the use of their data? : Yes
- Provide further information :
- Will payment be made to research participants? :No
- Provide further information :
- Is the research Health Research as defined in the Health Research Regulations? : No
- Does the research require a Consent Declaration form as defined by the Health Research Regulations 2018 and amendments 2021? :
- Which of the following best describes the research Secondary Data Source(s)? :
- Could any of the research data directly identify any participant? : No
- Could any of the research data indirectly identify any participant? :No
- Do you process Personal Data for study administration purposes? e.g. contacting individuals : No
- If yes please outline how you will store the information, keep it confidential, and destroy it in line with the data retention policy. :
- Which of the following best describes the general characteristics of the target population? :Adults currently not at risk of vulnerability
- Do any of the following further describe the characteristics of the target population? :
- Please describe :
- List the inclusion/ exclusion criteria for selection of research participants : Virtual Record Treasury of Ireland Domain Experts.

Research Sources and Sites

Name	Req
Google forms	NOT-APPLICABLE

Outline of research methods

No data available in this section.

-
- Does your research use any of the following methods exclusively? : Information, documents or data which are in the public domain

- Does the research include an intervention? : No
 - Please select which of the following best describes the intervention :
 - Other :
 - Outline the intervention :
 - Outline the method of analysis (Word limit :100 words) : The evaluation employs a Likert scale (1-5) to systematically assess each generated competency question across three key criteria: Relevance (How well does the question align with the purpose of competency questions?), Clarity (Is the question phrased in a clear, unambiguous, and grammatically correct way that reflects its intended meaning?), and Answerability (Can the knowledge graph reasonably provide an answer to the question?). Mean scores and standard deviations will be analyzed to identify patterns in question quality, while inter-rater reliability will quantify evaluator agreement. Additionally, qualitative feedback will undergo thematic analysis to detect potential patterns.
 - What is the approximate size of the target population? : 6
 - What is the proposed sample size -how many participants are involved in the study? : 4
 - What is the justification for the sample size? : The potential number of individuals with expert domain knowledge of the Virtual Record Treasury of Ireland project.
-
-

Risk

Risk , benefit, harm to researcher

- What setting/s will be used for data collection : Place of convenience for participants
- Other elaborate :

Type	Other Type	Impact	Probability	Mitigation
None				

Risk , benefit, harm to site environment or Society

Type	Impact	Probability	Mitigation
None			

Risk , benefit, harm to participants

- Do you or any of the research team have any dependant relationship to the participants? : : No

- Detail the role of the researcher/s, the relationship of the researcher/s to the participants normally, and how you will mitigate against coercion of participants to participate in the research or in influencing their responses :

Type	Other Type	Impact	Probability	Mitigation
None				

Participant Benefits

- Is it foreseeable that participants could potentially reveal information that you have a legal obligation to disclose (e.g. child protection policy, malpractice, etc.) :No

- What information may be disclosed, why, and to whom. :

- Outline any direct benefits to research participants of participation : To progress research related to a framework which can be used in their domain.

Conflict of Interest

- Are you aware of any Conflict of Interest from the PI or any collaborator, processor or other person involved in the conduct of the project, that could arise in the course of the project? : No

- Give details of the Conflict of Interest and what mitigation measures are in place :

Consent

Consent

- Do your participants require support to give Consent : No

- How will Assent be obtained and by whom :

- Provide Further Information :

- Do you require Assent or Proxy Consent from participants e.g. because of their vulnerability : No

- Provide Further Information :

- Are you required to have Garda Clearance : No

- What is the time interval between giving information and securing Consent : 7 or More Days

- Provide Further Information :

- Describe how you will inform participants about the use of their Personal Data : Participants will be made aware prior to the commencement of the experiment that their anonymised data may be used in publications related to the work as direct quotations. No identifiable information will be stored,

therefore, no identifying information will be contained in the publications. This information will be provided in the informed consent.

- Describe how participants can withdraw their Consent and/or their data : The study is fully anonymous, therefore, the confirmation page will state that no personal data has been recorded and that the study is fully anonymous.

Funding

Funding

- Insert RPAMS number if applicable and available :

- Outline sources of funding, list names of all confirmed sources of funding or support (including in-kind benefit). For all, state if they are industry/commercial, state/public, philanthropic/charitable, other : The research conducted was funded by the Irish Research Council under project ID GOIPG/2021/1165 and supported by the Research Ireland ADAPT Centre for Digital Content Technology (grant number 13/RC/2106 P2).

- Please specify any funder specific requirements or obligations which should be brought to the attention of the ethics committee and or Trinity Innovation & Enterprise : There are no such requirements.

- Will the results of the project be used or disclosed for commercial purposes : No

- Please clarify which party shall have the commercialisation and/or intellectual property rights : The Irish Research Council and the ADAPT Research Centre.

Conflict of Interest

- Are you aware of any possible Conflict of Interest arising from the funding or commercialisation of this project : : No

- Give details of the Conflict of Interest and what mitigation measures are in place :

- It there likely to be possible Conflict of Interest between the funders of the project and the aims and results of the project : No

- Give details of the Conflict of Interest and what mitigation measures are in place :

Sampling & Recruitment

Sampling and Recruitment

- Outline the sampling method : VRTI domain experts will be asked to evaluate the questions.
 - Describe the time commitment of the participant : The evaluation should take no more than 30 minutes to complete.
 - Will the research require/use a gatekeeper :No
 - Outline the position/role of the gatekeeper within the organisation :
 - Detail the role of the gatekeeper in the research :
 - Is there a dependant relationship between the gatekeeper and the participants :
 - Outline how this is going to be managed to mitigate against the dependencies :
 - Please give a detailed step by step description of how participants will be recruited and append the recruitment material (Word limit: 100 words). : The participants will be emailed with a link to the Google form directly.
-

A.4.4 User Evaluation 3: Ethics Application

Applicant & Collaborators

Applicant Details

- Applicant Name : Claire Mcnamara
- Is applicant applying as a member of staff or a student? : Student
- School / Department : School of Computer Science & Statistics - Computer Science
- Role on the Project : Principal Investigator
- Primary Employer (if not TCD) :
- Other affiliations (if applicable) :
- Course (for student applicant only) : PhD
- Part time / full time : Full-Time
- I have read and understood Trinity policy on Good Research Practice : Yes
- I have read and completed the Integrity module : Yes
- I have read and completed the Data Protection Training module within the last 24 months : No

Trinity Collaborators

User	Role
Declan O'Sullivan	Primary Supervisor
Lucy Hederman	Trinity Co supervisor

Other Collaborators

No data available in this section.

Project Details

Main Project Details

- Title of research :Tús Maith - Domain Expert Evaluation
- Data Collection Start Date : 01/12/2025
- Data Collection End Date : 25/12/2025
- Project End Date : 28/02/2026 12:00:00 AM
- Does the research involve :Humans (or their data)
- Does this Animal research involve :
- Is this AREC research a first application or an amendment :
- Is the research WHOLLY (i) an analysis of statutes/legal provisions AND/OR (ii) legal judgments/cases which have been made public by a judicial process? :
- Trinity Researchers will have access ONLY to aggregate data or data that otherwise fall outside of the remit of the GDPR. :
- Could the research have detrimental legal, economic or social consequences for either the participant or their establishments : No
- Intentions of the research: Does the research : None of the above
- State research aim(s) and objective(s), research question or hypothesis : See 2.1.11 for aim/objective and acronyms. Research question (RQ): To what extent does the TMv1 workflow aid domain experts in the efficient creation of high-quality, publishable CuQs for the purpose of Tús Maith? Sub-RQ1: How usable and efficient do domain experts find the TMv1 workflow? Sub-RQ2: To what extent does the provision of provenance and model selection transparency of TMv1 promote domain expert trust in LLM-generated content? Sub-RQ3: To what extent do the domain experts believe the CuQs produced align with the purpose of Tús Maith?
- Lay Summary: including background/ rationale/ justification, research approach, study design (exclude detail of measurement instruments and intervention and analysis (if applicable) (Word limit: 250 words) : The Virtual Record Treasury of Ireland (VRTI) Knowledge Graph (KG) is a large digital reconstruction of Irish historical records that were lost in the 1922 destruction of the Public Record Office of Ireland. It brings together rich archival data from multiple sources but, because of its complexity, can be difficult for non-technical users to explore or interpret. Tús Maith version 1 (TMv1) has been developed to address this challenge by supporting domain experts (for example, historians) in generating curated questions and answers (CuQs) based on the current contents of the KG. These CuQs are intended to act as exploratory starting points that help non-technical, non-domain-familiar people discover meaningful areas of the collection. The purpose of this evaluation is to understand to what extent the TMv1 workflow supports domain experts in producing such questions, and whether the resulting CuQs are of a quality that experts would consider suitable for publication on the public VRTI Explorer (an interface for the VRTI KG). The evaluation involves recruiting a small number of expert participants who will use TMv1 to complete the three-stage curation workflow and provide feedback through a short think-aloud exercise and follow-up questionnaire. Findings will inform refinements to the workflow and contribute to broader research on making complex KGs more accessible to non-technical users.
- Identify all countries where data is collected or processed : IE
- Does the project involve :Human participants and /or their data and no biological samples

- Is the project funded : Yes

Details on Human Participants and their Data

- Does your research involve Mixed Methods : No
 - Does the project use data from :Primary sources only
 - Will you obtain Consent from participants (or in the case of children or adults at risk of vulnerability from a parent / legal guardian) for their participation and/ or the use of their data? : Yes
 - Provide further information :
 - Will payment be made to research participants? :No
 - Provide further information :
 - Is the research Health Research as defined in the Health Research Regulations? : No
 - Does the research require a Consent Declaration form as defined by the Health Research Regulations 2018 and amendments 2021? :
 - Which of the following best describes the research Secondary Data Source(s)? :
 - Could any of the research data directly identify any participant? : No
 - Could any of the research data indirectly identify any participant? :Yes
 - Do you process Personal Data for study administration purposes? e.g. contacting individuals : Yes
 - If yes please outline how you will store the information, keep it confidential, and destroy it in line with the data retention policy. : The only personal data necessary to be known for the evaluation is that the participant is a domain experts in history with familiarity with the VRTI project and KG. There will be voice recordings taken as part of the think-aloud protocol. These recordings will be coded with a participant number that aligns with their questionnaire response. Once the recordings have been transcribed by the principal researcher, they will be deleted (maximum 2 months post recording). The recordings will be captured using the iPhone voice memo app and uploaded to Google drive under a TCD email address. Once the recordings have been uploaded, they will be deleted off of the iPhone. The questionnaire responses and transcriptions of the voice recordings will be stored in the same Google drive account.
 - Which of the following best describes the general characteristics of the target population? :Adults currently not at risk of vulnerability
 - Do any of the following further describe the characteristics of the target population? :
 - Please describe :
 - List the inclusion/ exclusion criteria for selection of research participants : Inclusion criteria: Historical domain experts familiar with the Virtual Record Treasury of Ireland project and KG.
-

Research Sources and Sites

Name	Req
------	-----

Google forms	NOT-APPLICABLE
Nvivo	LICENSE
iPhone Voice Memo App	NOT-APPLICABLE

Outline of research methods

No data available in this section.

- Does your research use any of the following methods exclusively? : Information, documents or data which are in the public domain
- Does the research include an intervention? : No
- Please select which of the following best describes the intervention :
- Other :
- Outline the intervention :
- Outline the method of analysis (Word limit :100 words) : Participant responses will be analysed using mixed methods. Quantitative data from the questionnaire will be summarised using descriptive statistics (means and standard deviations) to assess perceived workload and trust in the TMv1 workflow. Behavioural metrics (time-on-task, clicking on provenance-related buttons) will be triangulated with these results. Verbal data from the think-aloud sessions and open-text responses will undergo thematic analysis to identify patterns related to perceived usefulness, trust, and workflow efficiency. This integrated analysis will determine the extent to which the TMv1 workflow supports domain experts in creating publishable CuQs.
- What is the approximate size of the target population? : 8
- What is the proposed sample size -how many participants are involved in the study? : 5
- What is the justification for the sample size? : The number of individuals with sufficient knowledge of the contents of the VRTI KG.

Risk

Risk , benefit, harm to researcher

- What setting/s will be used for data collection : Place of convenience for participants
- Other elaborate :

Type	Other Type	Impact	Probability	Mitigation
None				

Risk , benefit, harm to site environment or Society

Type	Impact	Probability	Mitigation
None			

Risk , benefit, harm to participants

- Do you or any of the research team have any dependant relationship to the participants? : : No
- Detail the role of the researcher/s, the relationship of the researcher/s to the participants normally, and how you will mitigate against coercion of participants to participate in the research or in influencing their responses :

Type	Other Type	Impact	Probability	Mitigation
None				

Participant Benefits

- Is it foreseeable that participants could potentially reveal information that you have a legal obligation to disclose (e.g. child protection policy, malpractice, etc.) :No
 - What information may be disclosed, why, and to whom. :
 - Outline any direct benefits to research participants of participation : The TMv1 interface and workflow has been designed to aid the participants in generating meaningful starting points into an unfamiliar KG. They are contributing to research that aims to make the complex KG more accessible and easier to navigate for the public.
-

Conflict of Interest

- Are you aware of any Conflict of Interest from the PI or any collaborator, processor or other person involved in the conduct of the project, that could arise in the course of the project? : No
 - Give details of the Conflict of Interest and what mitigation measures are in place :
-

Consent

Consent

- Do your participants require support to give Consent : No
- How will Assent be obtained and by whom :

- Provide Further Information :
- Do you require Assent or Proxy Consent from participants e.g. because of their vulnerability : No
- Provide Further Information :
- Are you required to have Garda Clearance : No
- What is the time interval between giving information and securing Consent : 7 or More Days
- Provide Further Information :
- Describe how you will inform participants about the use of their Personal Data : Participants will be informed via the participant information leaflet.
- Describe how participants can withdraw their Consent and/or their data : Each participant will have a number assigned to them. This will be assigned using their Google form response and made known to the participant. If they wish to withdraw, they can email the principal researcher providing their number and the data associated with that number will be deleted.

Funding

Funding

- Insert RPAMS number if applicable and available :
- Outline sources of funding, list names of all confirmed sources of funding or support (including in-kind benefit). For all, state if they are industry/commercial, state/public, philanthropic/charitable, other : The research conducted was funded by the Irish Research Council under project ID GOIPG/2021/1165, it was also funded by the Trinity College Dublin School of Computer Science and Statistics, and supported by the Research Ireland ADAPT Centre for Digital Content Technology (grant number 13/RC/2106 P2).
- Please specify any funder specific requirements or obligations which should be brought to the attention of the ethics committee and or Trinity Innovation & Enterprise : There are no such requirements.
- Will the results of the project be used or disclosed for commercial purposes : No
- Please clarify which party shall have the commercialisation and/or intellectual property rights : The Irish Research Council, the ADAPT Research Centre, and the TCD School of Computer Science and Statistics.

Conflict of Interest

- Are you aware of any possible Conflict of Interest arising from the funding or commercialisation of this project : : No
- Give details of the Conflict of Interest and what mitigation measures are in place :

- It there likely to be possible Conflict of Interest between the funders of the project and the aims and results of the project : No

- Give details of the Conflict of Interest and what mitigation measures are in place :

Sampling & Recruitment

Sampling and Recruitment

- Outline the sampling method : The population for consideration will be historians with familiarity with the VRTI project and KG. They will be contacted directly by the principal researcher and asked to partake in the evaluation.
 - Describe the time commitment of the participant : 30 minutes.
 - Will the research require/use a gatekeeper :No
 - Outline the position/role of the gatekeeper within the organisation :
 - Detail the role of the gatekeeper in the research :
 - Is there a dependant relationship between the gatekeeper and the participants :
 - Outline how this is going to be managed to mitigate against the dependencies :
 - Please give a detailed step by step description of how participants will be recruited and append the recruitment material (Word limit: 100 words). : The principal researcher will email the participants directly attaching the participant information leaflet to ask them to partake in the evaluation. The email addresses will be obtained through the VRTI project. The evaluation itself will take place at the participants place of work at a day/time of convenience to the participant.
-

Data Protection

Opening Questions

- Have all Trinity researchers (staff and students) in this project completed the College Data Protection GDPR training module? * :No
 - Are all Trinity Staff and Trinity Students familiar with the Trinity College Personal Data Breach Procedural Guidelines? : Yes
-

Data Protection Assesement

- What is Trinity's role in the research? : Controller
- How many participants' Personal Data is being processed in this project? : <100
- List all types of Personal Data (including any special category or sensitive Personal Data) that you will process during the lifecycle of the project? : Google form questionnaire. Audio recording. Audio transcriptions.
- Does the research involve processing of special category data or data relating to criminal convictions and/or offences (sensitive personal data)? : No
- Is the Personal Data shared outside the research team with any other units within Trinity College? :No
- Name these units :
- Is this data shared with any third party outside of Trinity ? : No
- Provide names of these organisations and detail what Personal Data will be shared with them and why. :
- Describe what IT due diligence you have carried out on these organisations. : The audio recordings will immediately be uploaded to a Google drive account associated with a TCD email address. Once uploaded, they will be deleted from the iPhone.
- Provide a general description of the security measures in place to keep Personal Data secure for each system, platform and application you will use for access, storage, and transfer, including but not limited to multi factor authentication, use of passwords, use of VPN, device encryption, vendor ISO certification, anti-virus used, use of secure file transfers such as HEA net, and detail on how data is backed up etc. : The Google drive where the recordings will be stored until transcribed and, what houses the Google form, is associated with a TCD email address. This account is protected via 2 factor authentication.
- Is the data transferred outside of the UK & EEA? If yes, you must contact the Data Protection Office : No
- Detail how long Personal Data will be retained in an identifiable or coded format. : Once the voice recording has been transcribed, it will be deleted. This will occur within a maximum 2 month period after the recording has been obtained (February 2026). The questionnaire responses will remain until after the principal researcher has carried out their Viva (May 2026) and the research has been published after which they will be deleted.

Processing Risk

Type	Impact	Probabilty	Mitigation
HACKING OF THE GOOGLE DRIVE ACCOUNT.	Low	Low	THERE IS 2 FACTOR AUTHENTICATION ON THE ACCOUNT.

Closing Section

- Include any additional information in respect of the research which may be relevant * :N/A

A.5 LLM Prompts Used

A.5.1 LLM Experiment 1: Prompt

Listing A.1: Competency question template prompt with condition annotations.

```
# Template-style competency question prompt.
# Modular components annotated with experimental conditions.
# -----

{INCLUDE_ROLE? You are a historical knowledge graph exploration assistant. Your
  task is to generate natural language, template-style competency questions
  that ordinary users might ask when exploring a historical knowledge graph.}
[Condition(s):  Prompt 2,
+Role]

I am providing you with a set of knowledge graph classes, properties
{INCLUDE_RELATIONSHIPS? , and relationships}. Using only the provided
information, generate 20 template-style competency questions that could be
asked about this ontology. [Condition(s):  All; relationships only in Subset
and CIDOC CRM Relationships, +Relationships]

A competency question in this context is defined as a natural language question
that helps convey the scope and potential of a KG to users unfamiliar with
its structure with the aim of supporting exploration.

{INCLUDE_EXAMPLES? Examples of the type of questions I am looking for are:
  "Where was a person born?", "When did a person die?", "What type of place is
  this?".} [Condition(s):  Few Shot,
+Examples]

Phrase questions in natural, user-friendly language, avoiding technical or
ontology-specific terms e.g. use "was born" instead of "birth event".

Use template-style phrasing: general questions with placeholders like "a
person", "a place" etc.

Focus on creating general, reusable question templates that can later be filled
with specific instance data.

Only use the given classes, properties {INCLUDE_RELATIONSHIPS? , and
relationships}; do not invent additional concepts. [Condition(s):  All;
relationships only in Subset and CIDOC CRM Relationships, +Relationships]

You may combine classes and properties if it helps express a useful question.

Phrase each competency question naturally, as if it were intended for a user
exploring the knowledge graph.
```

A.5.2 LLM Experiment 2: Prompts Used

A.5.2.1 Basic Provenance

Listing A.2: Basic Provenance.

```
### Basic Provenance Prompt
Your task is to fill in template questions and answers using data from a JSON
    object and track the source information for each answer.

INPUT
You will be provided with:
1. A list of template questions.
2. A JSON object containing information about multiple entities.

OUTPUT FORMAT:
You MUST output a single, valid JSON object with the following structure:
{
  "results": [
    {
      "question": "[The complete filled question]",
      "answer": "[The concise answer]",
      "provenance": {
        "source_entity_uri": "https://..."
      }
    }
  ]
}

RULES
1. Completeness: Only include questions you can definitively answer. If a
    required JSON field is missing or empty (""), omit the question entirely.
2. Question Content: Fill in the blanks in the template question using the main
    entity's name and other details.
3. Answer Content: The answer must be a concise fact, not a full sentence. Use
    the most specific, human-readable field available (e.g., prefer
    `DisplayBirthDate` over `BirthDateUpper`).
4. Provenance: For every answered question, you must record the
    `source_entity_uri`.
5. No External Knowledge: Use only the information provided in the input JSON.
6. Strict Format: Do not add any text, explanations, or code fences (e.g.,
    ```json) before or after the JSON object.
```

---

### A.5.2.2 Full Provenance

Listing A.3: Basic Provenance.

---

```
Full Provenance Prompt
```

Your task is to fill in template questions and answers using data from a JSON object and track the source information for each answer.

#### INPUT

You will be provided with:

1. A list of template questions.
2. A JSON object containing information about multiple entities.

#### OUTPUT FORMAT:

You MUST output a single, valid JSON object with the following structure:

```
{
 "results": [
 {
 "question": "[The complete filled question]",
 "answer": "[The concise answer]",
 "provenance": {
 "source_entity_uri": "https://...",
 "evidence": [
 {"json_key": "FieldName", "json_value": "FieldValue"}
]
 }
 }
]
}
```

#### RULES

1. Completeness: Only include questions you can definitively answer. If a required JSON field is missing or empty (""), omit the question entirely.
2. Question Content: Fill in the blanks in the template question using the main entity's name and other details.
3. Answer Content: The answer must be a concise fact, not a full sentence. Use the most specific, human-readable field available (e.g., prefer ``DisplayBirthDate`` over ``BirthDateUpper``).
4. Provenance: For every answered question, you must record the ``source_entity_uri``, all the ``json_key``'s used and all the ``json_value``'s used to create the answer.
5. No External Knowledge: Use only the information provided in the input JSON.
6. Strict Format: Do not add any text, explanations, or code fences (e.g., ````json`) before or after the JSON object.

---

### A.5.3 User Evaluation 3: Prompts Used

#### A.5.3.1 Stage 1: Prompt

---

Listing A.4: Stage 1: CQ Template Derivation.

---

### Stage 1, Generate CQ Templates Prompt

You are an expert ontology engineer tasked with communicating the scope of a knowledge graph to non-technical users.

I will provide a json object of classes, object properties, and data properties from an ontology of RDF TTL information.

Using only the information I give you, generate exactly 20 high-quality template competency questions in natural, everyday English.

A competency question in this context is defined as a natural language question that helps convey the scope and potential of a KG to users unfamiliar with its structure with the aim of supporting exploration.

These questions should:

- Sound like something a curious user would actually ask when exploring the knowledge graph
- Use simple, user-friendly wording (e.g., "was born in" instead of "hasBirthPlace", "works at" instead of "employedBy")
- Be phrased as reusable templates using generic placeholders ("a person", "an organisation", "a book", "some event", etc.)
- Never use technical ontology terms (no "rdf:type", "owl:ObjectProperty", URIs, prefixes, or camelCase property names)
- Be answerable using only the classes and properties provided
- Be able to be filled in using a single entity, multiple entities as the answer are allowed
- As a set of questions, convey the breadth and scope of the knowledge graph

Good examples:

"Where was a person born?"

"What era does a place belong to?"

Bad examples (avoid):

"What is the birthPlace of a E21:Person?"

Return your response strictly as valid JSON in this exact format (nothing else - no explanations, no markdown, no extra text, no backticks):

```
{
 "questions": [
 "Question one?",
 "Question two?",
 "...",
 "Question twenty?"
]
}
```

---

### A.5.3.2 Stage 2: Prompt

### Listing A.5: Stage 2: Basic Provenance.

---

#### ### Stage 2, Basic Provenance Prompt

Your task is to fill in template questions and answers using data from a JSON object and track the source information for each answer.

#### INPUT

You will be provided with:

1. A list of template questions.
2. A JSON object containing information about multiple entities.

#### OUTPUT FORMAT:

You MUST output a single, valid JSON object with the following structure:

```
{
 "results": [
 {
 "question": "[The complete filled question]",
 "answer": "[The concise answer]",
 "provenance": {
 "source_entity_uri": "https://..."
 }
 }
]
}
```

#### RULES

1. Completeness: Only include questions you can definitively answer by directly getting the answer from the source JSON or by reasoning over the JSON.
  2. Question Content: Fill in the blanks in the template question using the main entity's name and other details.
  3. Answer Content: Use the most precise information available. Use the most specific, human-readable field available (e.g., prefer `DisplayBirthDate` over `BirthDateUpper`). If you deem relevant, you may provide the answer as a concise sentence.
  4. Provenance: For every answered question, you must record the `source\_entity\_uri`.
  5. No External Knowledge: Use only the information provided in the input JSON.
  6. Strict Format: Do not add any text, explanations, or code fences (e.g., ```json) before or after the JSON object.
- 

## A.6 Entity Ranking Algorithm

### Core Functions

---

**Algorithm 1** Entity Ranking Main Workflow

---

```
0: Input:
0: ga_top_n_terms: Number of top search terms from GA
0: K: Top entities per term to return
0: person_slot_threshold: Minimum slot count for persons
0: place_slot_threshold: Minimum slot count for places
0: Output: Ranked entities CSV file
0: procedure MAIN
0: top_search_terms $\leftarrow$ GETTOPSEARCHTERMSFROMGA(ga_top_n_terms)
0: all_ranked_entities $\leftarrow$ $\emptyset$
0: for each term in top_search_terms do
0: persons $\leftarrow$ PROCESSENTITIES(term, "person", person_slot_threshold, K)
0: places $\leftarrow$ PROCESSENTITIES(term, "place", place_slot_threshold, K)
0: Append persons and places to all_ranked_entities
0: end for
0: Write all_ranked_entities to CSV file
0: end procedure=0
```

---

**Scoring Constants (Heuristics)**

$BASE\_EXACT = 1.00$   
 $BASE\_TOKEN\_EXACT = 0.95$   
 $BASE\_PREFIX = 0.70$   
 $BASE\_SUBSTRING = 0.60$   
 $BASE\_NONE = 0.30$   
 $NAME\_TOKEN\_BOOST = 0.12$   
 $PLACE\_BOOST\_COUNTY = 0.12$   
 $PLACE\_BOOST\_TOWNLAND = 0.06$   
 $PLACE\_BOOST\_BARONY\_PARISH = 0.04$   
 $PREFIX\_MIN\_LENGTH = 3$   
 $MAX\_SCORE = 1.6$

**SPARQL Query Templates (Simplified)**

**Coverage Query Person:** Returns persons with name containing search term, plus completeness indicators (birth/death dates/places, image, gender)

**Coverage Query Place:** Returns places with name containing search term, from present-day and modern graphs

**External Refs Queries:** Count owl:sameAs and crm:P71i\_is\_listed\_in links

**Degree Queries:** Count incoming/outgoing edges (excluding rdfs:label and

---

**Algorithm 2** Process Entities for a Search Term

---

```
0: procedure PROCESSENTITIES(search_term, entity_type, slot_threshold, K)
0: if entity_type = "person" then
0: df $\leftarrow$ Execute COVERAGEQUERYPERSON(search_term)
0: degree_query $\leftarrow$ DEGREE_QUERY_PERSON
0: else
0: df $\leftarrow$ Execute COVERAGEQUERYPLACE(search_term)
0: degree_query $\leftarrow$ DEGREE_QUERY_PLACE
0: end if
0: if df is empty then
0: return empty DataFrame
0: end if
0: Filter df where slotCount $\geq$ slot_threshold
0: for each row r in df do
0: r.matchScore $\leftarrow$ COMPUTEMATCHSCORE(r, search_term, entity_type)
0: end for
0: uris $\leftarrow$ Extract URIs from df
0: {Get external references}
0: if entity_type = "person" then
0: ext_df $\leftarrow$ Execute EXTERNALREFSQUERYPERSON(uris)
0: else
0: ext_present $\leftarrow$ Execute EXTERNALREFSQUERYPLACEPRESENT(uris)
0: ext_modern $\leftarrow$ Execute EXTERNALREFSQUERYPLACEMODERN(uris)
0: ext_df $\leftarrow$ Merge ext_present and ext_modern
0: end if
0: {Get graph degree}
0: degree_df $\leftarrow$ Execute degree_query(uris)
0: {Merge all data}
0: if ext_df not empty then
0: df $\leftarrow$ Merge(df, ext_df, on="Entity")
0: end if
0: if degree_df not empty then
0: df $\leftarrow$ Merge(df, degree_df, on="Entity")
0: end if
0: {Compute normalised scores}
0: norm_slot $\leftarrow$ df.slotCount / max(df.slotCount, 1)
0: norm_ext $\leftarrow$ df.externalRefs / max(df.externalRefs, 1)
0: norm_deg $\leftarrow$ df.degree / max(df.degree, 1)
0: coverageScore $\leftarrow$ $0.5 \times \text{norm_slot} + 0.25 \times \text{norm_ext} + 0.25 \times \text{norm_deg}$
0: finalScore $\leftarrow$ df.matchScore $\times$ ($0.5 + 0.5 \times \text{coverageScore}$)
0: Remove duplicates by URI, keeping highest finalScore
0: Sort by finalScore descending
0: return top K rows
0: end procedure=0
```

---

---

**Algorithm 3** Compute String Match Score

---

```
0: function COMPUTEMATCHSCORE(row, search_term, entity_type)
0: s_term $\leftarrow$ lowercase(search_term)
0: if entity_type = "person" then
0: label $\leftarrow$ lowercase(row.NameLabel)
0: first $\leftarrow$ lowercase(row.FirstNameLabel)
0: sur $\leftarrow$ lowercase(row.Surname)
0: else
0: label $\leftarrow$ lowercase(row.Name)
0: first $\leftarrow$ ""
0: sur $\leftarrow$ ""
0: end if
0: label_tokens $\leftarrow$ split(label)
0: first_tokens $\leftarrow$ split(first)
0: surname_tokens $\leftarrow$ split(sur)
0: search_tokens $\leftarrow$ split(s_term)
0: {Multi-token exact match for persons}
0: if entity_type = "person" and $|search_tokens| \geq 2$ then
0: if search_tokens[0] $\in$ first_tokens and search_tokens[-1] $\in$
0: surname_tokens then
0: base $\leftarrow$ BASE_EXACT
0: else
0: base $\leftarrow$ null
0: end if
0: else
0: base $\leftarrow$ null
0: end if
0: if base is null then
0: if label = s_term then
0: base $\leftarrow$ BASE_EXACT
0: else if $\exists t \in label_tokens : t = s_term$ then
0: base $\leftarrow$ BASE_TOKEN_EXACT
0: else if label_tokens[0] starts with s_term and $|s_term| \geq$
0: PREFIX_MIN_LENGTH then
0: base $\leftarrow$ BASE_PREFIX
0: else if s_term is substring of label then
0: base $\leftarrow$ BASE_SUBSTRING
0: else
0: base $\leftarrow$ BASE_NONE
0: end if
0: end if
0: {Apply boost}
0: boost $\leftarrow$ 0.0
0: if entity_type = "person" then
0: if s_term $\in$ first_tokens or s_term $\in$ surname_tokens then
0: boost $\leftarrow$ NAME_TOKEN_BOOST
0: end if
0: else
0: boost $\leftarrow$ PLACETYPEBOOST(row.PlaceType)
0: end if
0: score $\leftarrow$ base $\times$ (1.0 + boost)
0: return min(MAX_SCORE, score)
0: end function
```

---

**Algorithm 4** Place Type Boost Function

---

```
0: function PLACETYPEBOOST(placetype_iri)
0: if placetype_iri contains "County" then
0: return PLACE_BOOST_COUNTY
0: else if placetype_iri contains "Townland" then
0: return PLACE_BOOST_TOWNLAND
0: else if placetype_iri contains "Barony" or "Parish" then
0: return PLACE_BOOST_BARONY_PARISH
0: else
0: return 0.0
0: end if
0: end function=0
```

---

---

**Algorithm 5** Get Top Search Terms from Google Analytics

---

```
0: function GETTOPSEARCHTERMSFROMGA(limit)
0: Query GA4 for "search_performed" events with dimensions:
0: - eventName
0: - customEvent:place_name
0: - customEvent:people_name
0: Filter: eventName = "search_performed" and at least one name parameter is set
0: Order by eventCount descending
0: Limit: $limit \times overfetch_factor$
0: top_terms $\leftarrow \emptyset$
0: seen $\leftarrow \emptyset$
0: for each row in response do
0: raw_term $\leftarrow$ non-empty value from place_name or people_name
0: norm $\leftarrow$ normalise(unicodedata.normalize("NFKC", raw_term.strip().lower()))
0: if norm $\notin$ seen and ENTITYEXISTSINKG(norm) then
0: Add norm to top_terms
0: Add norm to seen
0: end if
0: if $|top_terms| \geq limit$ then
0: Break
0: end if
0: end for
0: return top_terms
0: end function=0
```

---

crm:P1\_is\_identified\_by for persons)

## Key Data Structures

- **DataFrame**: Tabular structure containing entity data and scores
- **Entity**: URI identifying a person or place in the knowledge graph
- **Slot Count**: Sum of boolean indicators for entity completeness
- **Normalised Scores**: Values scaled to  $[0,1]$  range for combination

## A.7 Normalisation Rules

- `normalise_text` - generic text cleaner.
- `normalise_answer` - answer pre-processor (handles semicolon truncation).
- `canonicalise_answer` - core answer canonicalisation.
- `normalise_name` - person name normalisation using `nameparser`.
- `tokenise_set` - tokenisation for Jaccard similarity.

### A.7.1 Generic Text Normalisation (`normalise_text`)

1. If input is NaN or `pd.isna`, return empty string.
2. Convert to string.
3. Apply Unicode normalisation form **NFKC**.
4. Strip leading/trailing whitespace.
5. Replace any sequence of whitespace characters with a single space.
6. Remove all occurrences of: " ' ( ) [ ] \* ‘ (double quote, single quote, parentheses, square brackets, asterisk, backtick).
7. Replace the ampersand & with the word "and".

### A.7.2 Answer Normalisation

Answer strings undergo two phases.

#### A.7.2.1 Pre-processing (`normalise_answer`)

1. Convert to string, strip outer whitespace.
2. If the string contains a semicolon and the **last** semicolon is followed by at least one non-whitespace character, discard all text **before** that semicolon; keep only the substring after it.
3. Pass the result to `canonicalise_answer`.

#### A.7.2.2 Core Canonicalisation (`canonicalise_answer`)

1. Apply NFKC normalisation.
2. `_strip_noise`: strip leading/trailing spaces, commas, and periods.
3. Collapse multiple spaces into one.
4. **Date detection and conversion:**
  - If the string matches `[\d]{1,4}([-\/][\d]{1,2}){1,2}` (optionally with trailing text), attempt date parsing.
  - If the first numeric part  $> 12$ , try day-first parsing (DD/MM/YYYY).
  - Otherwise try month-first (MM/DD/YYYY), then day-first.
  - On success, return ISO date YYYY-MM-DD.
5. **Non-date cleaning:**
  - Convert the entire string to **lowercase**.
  - Replace sequences of two or more hyphens/dashes (`-`, `-`, `-`, `_`) with a single hyphen.
  - Replace any dash surrounded by optional whitespace with `-` (space-hyphen-space).
  - Remove all double quotes (`"`) and single quotes (`'`).
  - Collapse multiple spaces.
  - Insert a space between a lowercase letter and an uppercase letter (e.g., `vonHaller`  $\rightarrow$  `von Haller`).
6. Return the final cleaned string.

#### A.7.3 Person Name Normalisation (`normalise_name`)

Used to compare entity names extracted from questions with stored labels.

1. Parse the name using `nameparser.HumanName`.
2. Extract the `full_name` property (a cleaned, canonical full name).
3. If the parser returns an empty string, fall back to `normalise_text` on the original string.
4. Replace the original name with the parsed full name.
5. Apply `normalise_text` to the result.
6. During comparison, this string is **lowercased** (explicitly with `.lower()`).

#### A.7.4 Tokenisation for Jaccard Similarity (tokenise\_set)

Used when substring matching fails.

1. Apply `normalise_text` to the input.
2. Split the string using the regular expression `\W+` (any sequence of non-word characters).
3. Collect the resulting tokens into a set.

The Jaccard similarity is computed as:

$$J = \frac{|\text{gt\_tokens} \cap \text{gen\_tokens}|}{|\text{gt\_tokens} \cup \text{gen\_tokens}|}$$

A match is considered **acceptable** if  $J \geq 0.6$ .

### A.8 SPARQL Queries

#### A.8.1 Stage 1: SPARQL Query, Generate Classes and Properties and Relationships

Listing A.6: Stage 1: SPARQL.

---

```
Stage 1 SPARQL Query
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?subjectClass ?subjectClassLabel ?property ?propertyLabel
 ?objectClass ?objectClassLabel
WHERE {
 # Specify all graphs using VALUES and GRAPH with a variable
 VALUES ?graph {
 <https://kg.virtualtreasury.ie/graph/present-day-places-v1>
 <https://kg.virtualtreasury.ie/graph/modern-places-v1>
 <https://kg.virtualtreasury.ie/graph/DIB-v1>
 }
 GRAPH ?graph {
 ?s ?property ?o .

 # Get subject type and its label
 OPTIONAL {
 ?s rdf:type ?subjectClass .
 FILTER(ISIRI(?subjectClass))
 OPTIONAL {
 ?subjectClass rdfs:label ?subjectClassLabel
 FILTER(LANG(?subjectClassLabel) = "en")
 }
 }
 }
}
```

```

Get object type (if it's a resource) and its label
OPTIONAL {
 ?o rdf:type ?objectClass .
 FILTER(ISIRI(?objectClass))
 OPTIONAL {
 ?objectClass rdfs:label ?objectClassLabel
 FILTER(LANG(?objectClassLabel) = "en")
 }
}

Get property label
OPTIONAL {
 ?property rdfs:label ?propertyLabel
 FILTER(LANG(?propertyLabel) = "en")
}
}
ORDER BY ?subjectClass ?property

```

---

### A.8.2 Stage 2a: SPARQL Query, Determine Existence of Entities that Match a Search Term

Listing A.7: Stage 2a: SPARQL.

```

Stage 2a SPARQL Query
ASK {
 ?s rdfs:label ?label .
 FILTER(CONTAINS(LCASE(?label), "{term}"))
}

```

---

### A.8.3 Stage 2b: SPARQL Query, Retrieve the Necessary Information about a Given Entity in Order to Rank it

Listing A.8: Stage 2b: Coverage Query Person SPARQL

```

PREFIX crm: <http://erlangen-crm.org/current/>
PREFIX vrti: <https://www.w3id.org/virtual-treasury/ontology#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT DISTINCT
 ?Entity ?NameLabel ?Century ?FirstNameLabel ?Surname
 (BOUND(?BirthDateLower) AS ?hasBirthDate)
 (BOUND(?BirthPlaceLabel) AS ?hasBirthPlace)
 (BOUND(?DeathDateLower) AS ?hasDeathDate)
 (BOUND(?DeathPlaceLabel) AS ?hasDeathPlace)

```

```

(BOUND(?Image) AS ?hasImage)
(BOUND(?Gender) AS ?hasGender)
((xsd:integer(BOUND(?BirthDateLower)) +
 xsd:integer(BOUND(?BirthPlaceLabel)) +
 xsd:integer(BOUND(?DeathDateLower)) +
 xsd:integer(BOUND(?DeathPlaceLabel)) +
 xsd:integer(BOUND(?Image)) +
 xsd:integer(BOUND(?Gender))) AS ?slotCount)
WHERE {
 GRAPH <https://kg.virtualtreasury.ie/graph/DIB-v1> {
 ?Entity a crm:E21_Person ;
 vrti:century ?Century ;
 crm:P1_is_identified_by ?FirstName ;
 crm:P1_is_identified_by ?SurnameResource .

 ?FirstName crm:P2_has_type vrti:Forename ;
 rdfs:label ?FirstNameLabel .
 ?SurnameResource crm:P2_has_type vrti:Surname ;
 rdfs:label ?Surname .

 BIND(CONCAT(?FirstNameLabel, " ", ?Surname) AS ?NameLabel)

 OPTIONAL { ?Entity crm:P2_has_type ?Gender }
 OPTIONAL { ?Entity crm:P62i_is_depicted_by ?Image }

 OPTIONAL {
 ?BirthEvent a crm:E67_Birth ;
 crm:P98_brought_into_life ?Entity ;
 crm:P4_has_time-span ?BirthDateResource .
 ?BirthDateResource crm:P81a_end_of_the_begin ?BirthDateLower .
 OPTIONAL { ?BirthEvent crm:P7_took_place_at/rdfs:label ?BirthPlaceLabel
 FILTER(LANG(?BirthPlaceLabel)="en") }
 }

 OPTIONAL {
 ?DeathEvent a crm:E69_Death ;
 crm:P93_took_out_of_existence ?Entity ;
 crm:P4_has_time-span ?DeathDateResource .
 ?DeathDateResource crm:P81a_end_of_the_begin ?DeathDateLower .
 OPTIONAL { ?DeathEvent crm:P7_took_place_at/rdfs:label ?DeathPlaceLabel
 FILTER(LANG(?DeathPlaceLabel)="en") }
 }

 FILTER(CONTAINS(LCASE(?NameLabel), LCASE("%s")))
 }
}

```

---

Listing A.9: Stage 2b: Coverage Query Place SPARQL

---

```

PREFIX crm: <http://erlangen-crm.org/current/>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX vrtivocab: <https://www.w3id.org/virtual-treasury/vocabulary#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT DISTINCT
 ?Entity ?Name ?PlaceType ?TimePeriod
 (BOUND(?CountyLabel) AS ?hasCounty)
 (BOUND(?Centroid) AS ?hasCentroid)
 (BOUND(?Boundary) AS ?hasBoundary)
 ((xsd:integer(BOUND(?CountyLabel)) +
 xsd:integer(BOUND(?Centroid)) +
 xsd:integer(BOUND(?Boundary))) AS ?slotCount)
WHERE {
 {
 GRAPH <https://kg.virtualtreasury.ie/graph/present-day-places-v1> {
 ?Entity a crm:E53_Place ;
 rdfs:label ?Name ;
 crm:P2_has_type ?PlaceType .
 FILTER(LANG(?Name) = "en")
 FILTER(CONTAINS(LCASE(?Name), LCASE("%s")))
 BIND("Present-Day" AS ?TimePeriod)
 }
 }
 UNION
 {
 GRAPH <https://kg.virtualtreasury.ie/graph/modern-places-v1> {
 ?Entity a crm:E53_Place ;
 rdfs:label ?Name ;
 crm:P2_has_type ?PlaceType .
 FILTER(LANG(?Name) = "en")
 FILTER(STRSTARTS(STR(?PlaceType),
 "https://www.w3id.org/virtual-treasury/vocabulary#Modern-1750-1922-"))
 FILTER(CONTAINS(LCASE(?Name), LCASE("%s")))
 BIND("Modern-1750-1922" AS ?TimePeriod)
 }
 }

 OPTIONAL {
 ?Entity crm:P89_falls_within+ ?County .
 ?County crm:P2_has_type ?countyType ;
 rdfs:label ?CountyLabel .
 FILTER(LANG(?CountyLabel) = "en")
 FILTER(?countyType IN (vrtivocab:PresentDayCounty,
 vrtivocab:Modern-1750-1922-County))
 }
}

```

```

 OPTIONAL { ?Entity geo:hasCentroid/geo:asWKT ?Centroid }
 OPTIONAL { ?Entity geo:hasBoundingBox/geo:asWKT ?Boundary }
}

```

---

Listing A.10: Stage 2b: External References Query Person SPARQL

---

```

PREFIX crm: <http://erlangen-crm.org/current/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

SELECT ?Entity
 (COUNT(DISTINCT ?same) AS ?sameAsCount)
 (COUNT(DISTINCT ?listedIn) AS ?listedInCount)
 ((COUNT(DISTINCT ?same) + COUNT(DISTINCT ?listedIn)) AS ?externalRefs)
WHERE {
 GRAPH <https://kg.virtualtreasury.ie/graph/DIB-v1> {
 VALUES ?Entity { %s }
 OPTIONAL { ?Entity owl:sameAs ?same }
 OPTIONAL { ?Entity crm:P71i_is_listed_in ?listedIn }
 }
}
GROUP BY ?Entity

```

---

Listing A.11: Stage 2b: External References Query Place (Present) SPARQL

---

```

PREFIX crm: <http://erlangen-crm.org/current/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

SELECT ?Entity
 (COUNT(DISTINCT ?same) AS ?sameAsCount)
 (COUNT(DISTINCT ?listedIn) AS ?listedInCount)
WHERE {
 GRAPH <https://kg.virtualtreasury.ie/graph/present-day-places-v1> {
 VALUES ?Entity { %s }
 OPTIONAL { ?Entity owl:sameAs ?same }
 OPTIONAL { ?Entity crm:P71i_is_listed_in ?listedIn }
 }
}
GROUP BY ?Entity

```

---

Listing A.12: Stage 2b: External References Query Place (Modern) SPARQL

---

```

PREFIX crm: <http://erlangen-crm.org/current/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

SELECT ?Entity
 (COUNT(DISTINCT ?same) AS ?sameAsCount)
 (COUNT(DISTINCT ?listedIn) AS ?listedInCount)
WHERE {
 GRAPH <https://kg.virtualtreasury.ie/graph/modern-places-v1> {
 VALUES ?Entity { %s }
 }
}

```

```

 OPTIONAL { ?Entity owl:sameAs ?same }
 OPTIONAL { ?Entity crm:P71i_is_listed_in ?listedIn }
 }
}
GROUP BY ?Entity

```

---

#### Listing A.13: Stage 2b: Degree Query Person SPARQL

---

```

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX crm: <http://erlangen-crm.org/current/>

SELECT ?Entity
 (COUNT(?p) AS ?outDegree)
 (COUNT(?s) AS ?inDegree)
 ((COUNT(?p) + COUNT(?s)) AS ?degree)
WHERE {
 GRAPH <https://kg.virtualtreasury.ie/graph/DIB-v1> {
 VALUES ?Entity { %s }
 { ?Entity ?p ?o .
 FILTER(?p != rdfs:label && ?p != crm:P1_is_identified_by) }
 UNION
 { ?s ?q ?Entity .
 FILTER(?q != rdfs:label && ?q != crm:P1_is_identified_by) }
 }
}
GROUP BY ?Entity

```

---

#### Listing A.14: Stage 2b: Degree Query Place SPARQL

---

```

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?Entity
 (COUNT(DISTINCT ?p) AS ?outDegree)
 (COUNT(DISTINCT ?s) AS ?inDegree)
 ((COUNT(DISTINCT ?p) + COUNT(DISTINCT ?s)) AS ?degree)
WHERE {
 VALUES ?Entity { %s }

 # outgoing predicates across both graphs (exclude rdfs:label)
 {
 GRAPH <https://kg.virtualtreasury.ie/graph/present-day-places-v1> {
 ?Entity ?p ?o .
 FILTER(?p != rdfs:label)
 }
 }
 UNION
 {
 GRAPH <https://kg.virtualtreasury.ie/graph/modern-places-v1> {
 ?Entity ?p ?o .
 }
 }
}

```

```

 FILTER(?p != rdfs:label)
 }
}

incoming predicates across both graphs (exclude rdfs:label)
UNION
{
 GRAPH <https://kg.virtualtreasury.ie/graph/present-day-places-v1> {
 ?s ?q ?Entity .
 FILTER(?q != rdfs:label)
 }
}
UNION
{
 GRAPH <https://kg.virtualtreasury.ie/graph/modern-places-v1> {
 ?s ?q ?Entity .
 FILTER(?q != rdfs:label)
 }
}
}
GROUP BY ?Entity

```

---

#### A.8.4 Stage 2c: SPARQL Query, Retrieve the Semantic Neighbourhood of a Given Entity

Listing A.15: Stage 2c: Generate All Person Information SPARQL

---

```

PREFIX crm: <http://erlangen-crm.org/current/>
PREFIX vrti: <https://www.w3id.org/virtual-treasury/ontology#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT
 ?Person
 ?Gender
 ?Surname
 ?FullNameLabel
 ?FirstNameLabel
 ?DIB_Page
 ?TimePeriod
 ?DeathDateUpper
 ?DeathDateLower
 ?DisplayDeathDate
 ?BirthDateUpper
 ?BirthDateLower
 ?DisplayBirthDate
 ?DisplayFloruitLower

```

```

?DisplayFloruitUpper
?FloruitLowerYear
?FloruitUpperYear
(GROUP_CONCAT(DISTINCT ?SurnameLabel; separator = "---") AS ?SurnameLabels)
(GROUP_CONCAT(DISTINCT ?DeathPlace; separator = "---") AS ?DeathPlace)
(GROUP_CONCAT(DISTINCT ?BirthPlace; separator = "---") AS ?BirthPlace)
(GROUP_CONCAT(DISTINCT ?Image; separator = "---") AS ?Image)
WHERE {
 GRAPH <https://kg.virtualtreasury.ie/graph/DIB-v1> {
 VALUES ?Person { <%s> }
 ?Person a crm:E21_Person ;
 vrti:VRTI_ERA ?TimePeriod ;
 crm:P1_is_identified_by ?FullNameResource .
 OPTIONAL { ?Person crm:P2_has_type ?Gender }
 ?FullNameResource crm:P2_has_type vrti:Name ;
 rdfs:label ?FullNameLabel .
 OPTIONAL {
 ?Person crm:P1_is_identified_by ?FirstNameResource .
 ?FirstNameResource crm:P2_has_type vrti:Forename ;
 rdfs:label ?FirstNameLabel .
 OPTIONAL {
 ?FirstNameResource crm:P139_has_alternative_form ?ForenameVariant .
 ?ForenameVariant rdfs:label ?ForenameVariantLabel .
 }
 }
 }
 OPTIONAL {
 ?Person crm:P1_is_identified_by ?Surname .
 ?Surname crm:P2_has_type vrti:Surname ;
 rdfs:label ?SurnameLabel .
 OPTIONAL {
 ?Surname crm:P139_has_alternative_form ?SurnameVariant .
 ?SurnameVariant rdfs:label ?SurnameVariantLabel .
 }
 }
 BIND(CONCAT(COALESCE(?FirstNameLabel, ""), " ", COALESCE(?SurnameLabel, ""))
 AS ?Name)
 OPTIONAL { ?Person crm:P62i_is_depicted_by ?Image }
 OPTIONAL { ?Person owl:sameAs ?WikidataPage }
 OPTIONAL { ?Person crm:P71i_is_listed_in ?DIB_Page }
 BIND(STRAFTER(STR(?TimePeriod), "#") AS ?VrtiEraLabel)
 OPTIONAL {
 ?DeathEvent a crm:E69_Death ;
 crm:P93_took_out_of_existence ?Person .
 OPTIONAL {
 ?DeathEvent crm:P4_has_time-span ?DeathDateResource .
 ?DeathDateResource crm:P82b_end_of_the_end ?DeathDateUpper ;
 crm:P81a_end_of_the_begin ?DeathDateLower .
 }
 }
 BIND(

```

```

IF(
 BOUND(?DeathDateLower) && BOUND(?DeathDateUpper),
 IF(
 ?DeathDateLower = ?DeathDateUpper,
 CONCAT(STR(DAY(?DeathDateLower)), "-", STR(MONTH(?DeathDateLower)),
 "-", STR(YEAR(?DeathDateLower))),
 CONCAT(
 CONCAT(STR(DAY(?DeathDateLower)), "-",
 STR(MONTH(?DeathDateLower)), "-", STR(YEAR(?DeathDateLower))),
 " - ",
 CONCAT(STR(DAY(?DeathDateUpper)), "-",
 STR(MONTH(?DeathDateUpper)), "-", STR(YEAR(?DeathDateUpper)))
)
),
 IF(
 BOUND(?DeathDateLower),
 CONCAT(STR(DAY(?DeathDateLower)), "-", STR(MONTH(?DeathDateLower)),
 "-", STR(YEAR(?DeathDateLower))),
 IF(
 BOUND(?DeathDateUpper),
 CONCAT(STR(DAY(?DeathDateUpper)), "-",
 STR(MONTH(?DeathDateUpper)), "-", STR(YEAR(?DeathDateUpper))),
 ""
)
)
) AS ?DisplayDeathDate
)
}
OPTIONAL {
 ?DeathEvent crm:P7_took_place_at ?DeathPlaceResource .
 GRAPH <https://kg.virtualtreasury.ie/graph/present-day-places-v1> {
 ?DeathPlaceResource rdfs:label ?DeathPlaceLabel .
 FILTER(LANG(?DeathPlaceLabel) = "en")
 }
 BIND(CONCAT(STR(?DeathPlaceResource), ";", ?DeathPlaceLabel) AS
 ?DeathPlace)
}
}
OPTIONAL {
 ?BirthEvent a crm:E67_Birth ;
 crm:P98_brought_into_life ?Person .
 OPTIONAL {
 ?BirthEvent crm:P4_has_time-span ?BirthDateResource .
 ?BirthDateResource crm:P82b_end_of_the_end ?BirthDateUpper ;
 crm:P81a_end_of_the_begin ?BirthDateLower .
 BIND(
 IF(
 BOUND(?BirthDateLower) && BOUND(?BirthDateUpper),

```

```

IF(
 ?BirthDateLower = ?BirthDateUpper,
 CONCAT(STR(DAY(?BirthDateLower)), "-", STR(MONTH(?BirthDateLower)),
 "-", STR(YEAR(?BirthDateLower))),
 CONCAT(
 CONCAT(STR(DAY(?BirthDateLower)), "-",
 STR(MONTH(?BirthDateLower)), "-", STR(YEAR(?BirthDateLower))),
 " - ",
 CONCAT(STR(DAY(?BirthDateUpper)), "-",
 STR(MONTH(?BirthDateUpper)), "-", STR(YEAR(?BirthDateUpper)))
)
),
IF(
 BOUND(?BirthDateLower),
 CONCAT(STR(DAY(?BirthDateLower)), "-", STR(MONTH(?BirthDateLower)),
 "-", STR(YEAR(?BirthDateLower))),
 IF(
 BOUND(?BirthDateUpper),
 CONCAT(STR(DAY(?BirthDateUpper)), "-",
 STR(MONTH(?BirthDateUpper)), "-", STR(YEAR(?BirthDateUpper))),
 ""
)
)
) AS ?DisplayBirthDate
)
}
OPTIONAL {
 ?BirthEvent crm:P7_took_place_at ?BirthPlaceResource .
 GRAPH <https://kg.virtualltreasury.ie/graph/present-day-places-v1> {
 ?BirthPlaceResource rdfs:label ?BirthPlaceLabel .
 FILTER(LANG(?BirthPlaceLabel) = "en")
 }
 BIND(CONCAT(STR(?BirthPlaceResource), ";", ?BirthPlaceLabel) AS
 ?BirthPlace)
}
}
?FloruitResource crm:P2_has_type vrti:Floruit ;
 crm:P4_has_time-span ?FloruitDateResource ;
 crm:P12_occurred_in_the_presence_of ?Person .
?FloruitDateResource crm:P81a_end_of_the_begin ?FloruitLower ;
 crm:P82b_end_of_the_end ?FloruitUpper .
BIND(YEAR(?FloruitLower) AS ?FloruitLowerYear)
BIND(YEAR(?FloruitUpper) AS ?FloruitUpperYear)
BIND(CONCAT(STR(DAY(?FloruitLower)), "-", STR(MONTH(?FloruitLower)), "-",
 STR(YEAR(?FloruitLower))) AS ?DisplayFloruitLower)
BIND(CONCAT(STR(DAY(?FloruitUpper)), "-", STR(MONTH(?FloruitUpper)), "-",
 STR(YEAR(?FloruitUpper))) AS ?DisplayFloruitUpper)
}

```

}

---

Listing A.16: Stage 2c: Generate All Place Information SPARQL

---

```
PREFIX crm: <http://erlangen-crm.org/current/>
PREFIX vrti: <https://www.w3id.org/virtual-treasury/ontology#>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

SELECT DISTINCT
 ?CRM_Type
 ?EquivalentPlaceName
 ?Area
 ?LogainmLink
 ?FastLink
 ?TownlandsIELink
 ?DbpediaLink
 ?Centroid
 ?Point
 (GROUP_CONCAT(DISTINCT ?Name; separator = "---") AS ?Names)
 (GROUP_CONCAT(DISTINCT ?FallsWithinPlace; separator = "---") AS
 ?FallsWithinPlaces)
 (GROUP_CONCAT(DISTINCT ?FallsWithinPlaceType; separator = "---") AS
 ?FallsWithinPlaceTypes)
 (GROUP_CONCAT(DISTINCT ?FallsWithin; separator = "---") AS ?FallsWithinins)
 (GROUP_CONCAT(DISTINCT ?FallsWithinPlaceLabel; separator = "---") AS
 ?FallsWithinPlaceLabels)
 (GROUP_CONCAT(DISTINCT ?parentLabel; separator = "---") AS ?parentLabels)
WHERE {
 GRAPH <https://kg.virtualtreasury.ie/graph/present-day-places-v1> {
 VALUES ?Place { <%s> }
 ?Place a crm:E53_Place ;
 crm:P2_has_type ?CRM_Type .
 OPTIONAL {
 ?Place rdfs:label ?Name .
 FILTER(langMatches(lang(?Name), "en"))
 }
 OPTIONAL {
 ?Place crm:P89_falls_within ?FallsWithinPlace .
 ?FallsWithinPlace crm:P2_has_type ?FallsWithinPlaceType ;
 rdfs:label ?FallsWithinPlaceLabel .
 BIND(CONCAT(STR(?FallsWithinPlace), ";", ?FallsWithinPlaceLabel) AS
 ?FallsWithin)
 FILTER(langMatches(lang(?FallsWithinPlaceLabel), "en"))
 }
 OPTIONAL {
 ?Place crm:P89_falls_within+ ?parentPlace .
 ?parentPlace crm:P2_has_type ?parentType ;
```

```

 rdfs:label ?parentLabel .
 }
 GRAPH <https://kg.virtualtreasury.ie/graph/modern-places-v1> {
 OPTIONAL {
 ?EquivalentPlace crm:P189_approximates ?Place .
 ?EquivalentPlace rdfs:label ?EquivalentPlaceName .
 BIND(CONCAT(STR(?EquivalentPlace), "; ", STR(?EquivalentPlaceName)) AS
 ?EquivalentPlaceDetails)
 FILTER(langMatches(lang(?EquivalentPlaceName), "en"))
 }
 }
 OPTIONAL {
 ?Place geo:hasCentroid ?Centroid .
 ?Centroid geo:asWKT ?Point .
 }
 OPTIONAL { ?Place geo:hasMetricArea ?Area . }
 OPTIONAL {
 ?Place owl:sameAs ?LogainmLink .
 FILTER(STRSTARTS(STR(?LogainmLink), "http://data.logainm.ie/"))
 }
 OPTIONAL {
 ?Place owl:sameAs ?FastLink .
 FILTER(STRSTARTS(STR(?FastLink), "http://id.worldcat.org/fast/"))
 }
 OPTIONAL {
 ?Place crm:P71i_is_listed_in ?TownlandsIELink .
 FILTER(STRSTARTS(STR(?TownlandsIELink), "http://www.townlands.ie/"))
 }
 OPTIONAL {
 ?Place owl:sameAs ?DbpediaLink .
 FILTER(STRSTARTS(STR(?DbpediaLink), "http://dbpedia.org/resource/"))
 }
}

```

---

## A.8.5 LetterSampo Application: SPARQL Queries

### A.8.5.1 LetterSampo: Stage 1

Listing A.17: Stage 1: Retrieve Classes and Properties and Relation

```

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?subjectClass ?subjectClassLabel ?property ?propertyLabel
 ?objectClass ?objectClassLabel
WHERE {
 ?s ?property ?o .

```

```

Get subject type and its label
OPTIONAL {
 ?s rdf:type ?subjectClass .
 FILTER(ISIRI(?subjectClass))
 OPTIONAL {
 ?subjectClass rdfs:label ?subjectClassLabel
 FILTER(LANG(?subjectClassLabel) = "en")
 }
}

Get object type (if it's a resource) and its label
OPTIONAL {
 ?o rdf:type ?objectClass .
 FILTER(ISIRI(?objectClass))
 OPTIONAL {
 ?objectClass rdfs:label ?objectClassLabel
 FILTER(LANG(?objectClassLabel) = "en")
 }
}

Get property label
OPTIONAL {
 ?property rdfs:label ?propertyLabel
 FILTER(LANG(?propertyLabel) = "en")
}
}
ORDER BY ?subjectClass ?property

```

---

#### A.8.5.2 LetterSampo: Stage 2, Retrieve Seed Entities

Listing A.18: Stage 2: Top Places by Connections SPARQL

---

```

PREFIX crm: <http://www.cidoc-crm.org/cidoc-crm/>

SELECT ?place (COUNT(*) AS ?connections)
WHERE {
 ?place a crm:E53_Place .

 {
 ?place ?p ?o .
 }
 UNION
 {
 ?s ?p ?place .
 }
}
GROUP BY ?place

```

```
ORDER BY DESC(?connections)
LIMIT 10
```

---

#### Listing A.19: Stage 2: Top People by Connections SPARQL

---

```
PREFIX crm: <http://www.cidoc-crm.org/cidoc-crm/>
```

```
SELECT ?person (COUNT(*) AS ?connections)
WHERE {
 ?person a crm:E21_Person .

 {
 ?person ?p ?o .
 }
 UNION
 {
 ?s ?p ?person .
 }
}
GROUP BY ?person
ORDER BY DESC(?connections)
LIMIT 10
```

---

#### Listing A.20: Stage 2: Top Letters by Connections SPARQL

---

```
PREFIX ldf: <http://ldf.fi/schema/lssc/>
```

```
SELECT ?letter (COUNT(*) AS ?connections)
WHERE {
 ?letter a ldf:Letter .

 {
 ?letter ?p ?o .
 }
 UNION
 {
 ?s ?p ?letter .
 }
}
GROUP BY ?letter
ORDER BY DESC(?connections)
LIMIT 10
```

---

### A.8.5.3 LetterSampo: Stage 2, Retrieve Semantic Neighbourhood

#### Listing A.21: Stage 2: Retrieve Semantic Neighbourhood

---

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
```

PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

SELECT DISTINCT

  ?entity ?entityLabel ?entityType  
  ?predicate ?predicateLabel  
  ?value ?valueLabel ?valueType  
  ?direction

WHERE {

  VALUES ?entity { <\%s> }

  OPTIONAL { ?entity rdfs:label|skos:prefLabel ?entityLabel }

  OPTIONAL { ?entity a ?entityType }

{  
  ?entity ?predicate ?value .  
  BIND("outgoing" AS ?direction)  
}

UNION

{  
  ?value ?predicate ?entity .  
  BIND("incoming" AS ?direction)  
}

FILTER(?predicate NOT IN (rdf:type, rdfs:label, skos:prefLabel))

OPTIONAL { ?predicate rdfs:label|skos:prefLabel ?predicateLabel }

OPTIONAL { ?value a ?valueType }

OPTIONAL {  
  FILTER(isIRI(?value))

  OPTIONAL { ?value rdfs:label ?v11 }

  OPTIONAL { ?value skos:prefLabel ?v12 }

  OPTIONAL {  
    ?value ?p ?v2 .  
    ?v2 rdfs:label|skos:prefLabel ?v13 .  
    FILTER(?v2 != ?entity)  
  }

  BIND(COALESCE(?v11, ?v12, ?v13) AS ?valueLabel)

}

}

ORDER BY ?direction ?predicateLabel ?valueLabel

## Appendix B

# Future Work Extended

The following section consists of a potential future research direction to create multi-hop CuQAs.

### B.1 Augmenting TMv1 to Support Multi-Hop CuQA Generation

Domain expert feedback obtained during the evaluation of TMv1 indicated a desire for CuQAs that move beyond single-relation or attribute-centric frames toward questions that capture broader relational context. This section responds to this requirement by positioning the generation of structurally complex, multi-hop CuQAs as an extension of the existing CuQA lifecycle. Rather than introducing new interaction paradigms, the proposed approach augments the schema-driven, LLM-supported CQ template derivation process with explicit information derived from KG topology, enabling the systematic design of multi-relational CQ templates and the future formalisation of additional validity constraints required for their instantiation. The proposed approach outlined below is currently theoretical and therefore forms the basis for the future work. Furthermore, this section is not intended to be exhaustive of all potential approaches, but illustrative of a potential method of multi-hop CQ template generation using LLMs. It should therefore be read as a conceptual design exploration rather than a validated architectural requirement.

#### B.1.1 Design Motivation

While simple CuQAs are a necessary component of a CuQA set, as they explicitly communicate the basic scope and underlying ontology of a KG (see Chapter 5), qualitative feedback indicated a desire to extend TMv1 to support more structurally complex, multi-hop CuQAs. Although this possibility had been hypothesised during the design of TMv1, the empirical focus of the present work was on establishing the viability of the three-stage Tús Maith framework rather than on developing structurally complex CuQAs. In other words, the need to support the creation of multi-hop CuQAs emerged from empirical qualitative feedback and is not merely a theoretical embellishment of TMv1.

### B.1.2 KG Topology Informed CQ Template Generation

The current method of CQ template derivation in TMv1 does not take into account the structural and order-theoretic properties of a structured, directed schema graph underlying a KG, including connectedness, in- and out-degree of nodes, cardinality patterns, and their hierarchical nature. Meta-paths [167] provide a schema-level abstraction for representing multi-hop relational patterns over heterogeneous KGs. Unlike instance-level graph patterns which enumerate concrete sequences of entities and relations in the data, meta-paths specify typed sequences of class-property-class transitions that capture a general semantic pattern of connectivity within the ontology. CQs templates in the context of Tús Maith (i.e., CQ templates to be instantiated with entities to form exploratory starting points), can be considered a natural language representation of a subset of meta-paths as, while all CQ templates are structural templates, not all meta-paths are suitable to become CQ templates. In other words, meta-paths can be viewed as structural priors over the space of admissible single- and multi-hop CQ templates encoding which classes and relations may be composed into semantically coherent question frames independently of any particular entity selection.

While SPARQL permits traversal of graph edges in both forward and reverse directions through the use of inverse property paths, this capability reflects the expressive power of the query language rather than the epistemic structure of the ontology schema. In Tús Maith, directed ontology relations are treated as encoding asymmetric conceptual perspectives, particularly in event-centric models such as CIDOC CRM. The topology introduced here is not intended to characterise what is computationally retrievable, but rather what is presented as an admissible line of inquiry from a given entry class.

### B.1.3 Multi-Hop CQ Templates as a Subset of Meta-Paths

Much of the existing literature on meta-paths employs them as analytic or ranking instruments, for example, to compute structural similarity between entities [167], or to guide recommendation, clustering, and relevance scoring in heterogeneous networks [167, 168]. In such settings, meta-paths are evaluated over instance-level graphs to quantify how strongly two entities are connected under a given relational pattern. However, as pointed out by [169], identifying meta-paths for complex heterogeneous information networks (HINs) like KGs is a challenging process. Their proposed mechanism, SchemaWalk, for automatically discovering informative meta-paths in schema-complex HINs is presented as a way of supporting downstream tasks such as classification, link prediction, and node embedding. As such, meta-paths are treated as latent structural features whose value is determined by their contribution to model performance over instance data e.g., link prediction or representation learning.

In contrast, the role of meta-paths in this thesis is not analytic or predictive over instance data, but constructive and constraining over the design space of CQ templates. In Tús Maith, meta-paths are not learned from the data for the purpose of model optimisation; rather, they are treated as schema-level relational patterns that delimit the space of structurally admissible multi-hop CQ templates. While Liu et al. leverage meta-paths to

identify which connectivity motifs explain or predict instance level behaviour, this thesis proposes using them to delimit the space of meaningful multi-relational questions that can be presented to an LLM for multi-hop CQ template generation.

Constraining such meta-paths to only maintain CQ templates faithful to the directionality of the underlying ontology is hypothesised by the author of this thesis to be possible using Alexandroff topology [170]. This is grounded in the fact that the KG schema can be interpreted as inducing a preorder under reachability (reflexive:  $\forall x \in X, x \leq x$ , transitive: if  $x \leq y$  and  $y \leq z$ ,  $x \leq z$ ) and this preorder naturally induces an Alexandroff topology where the minimal open sets correspond exactly to the set of reachable nodes within the schema. In KG terms, reflexive closure represents the fact that every class is considered reachable from itself, even if there is no explicit loop edge and transitive closure refers to the fact that if it is possible to go from class  $A$  to class  $B$ , and from class  $B$  to class  $C$ , then  $A$  can be treated as related to  $C$  even if there is no direct edge between  $A$  and  $C$ .

In overlaying Alexandroff topology on the space of admissible meta-paths induced by schema preorder, it is possible to formally define the observable neighbourhood of a class within the schema graph. This constrains the generative space of the LLM such that any CQ template it produces is conditioned on structurally reachable regions of the ontology, rather than crossing into disjoint or epistemically inadmissible areas of the schema. Alexandroff topology is introduced not merely as a reachability formalism, but because it provides a principled notion of observable neighbourhoods under asymmetric conceptual relations, which is not easily captured through path enumeration alone. While these constraints could be implemented procedurally through schema traversal rules, Alexandroff topology provides one formal lens through which admissible neighbourhoods and directional reachability can be characterised.

More specifically, if Class  $A$  is connected to Class  $B$  (directly or via hops), it is possible to say  $A \leq B$ . Therefore, the minimal open set of Class  $A$  is the set of all reachable classes from  $A$  including  $A$ :

$$U_A = \{B \in X | A \leq B\}$$

Where  $X$  is the set of all Classes in the KG and the preorder  $\leq$  is defined by the directed edges as stated above.

In this formulation,  $U_A$  characterises the observable structural region of the ontology rooted in  $A$ . It therefore is possible to constrain meta-paths which can form multi-hop CQ templates to ones that are within the observable neighbourhood of  $A$ . More concretely, and more importantly for CuQAs generated using TMv1, this grounds multi-hop CQ templates in the ontology of the KG by constraining the generative space of the LLM to schema-reachable regions. Enforcing Alexandroff topology acts as a formal pre-generation filter over admissible meta-paths, thereby increasing the likelihood that generated multi-hop CQ templates are structurally answerable under the KG schema. An example from the VRTI KG can be defined as follows:

Consider the following directed relations:

E67\_Birth - P98\_brought\_into\_life  $\rightarrow$  E21\_Person

E67\_Birth - P7\_took\_place\_at  $\rightarrow$  E53\_Place

$E53\_Place - P89\_falls\_within \rightarrow E53\_Place$

A visualisation of these directed relations can be seen in Figure B.1 demonstrating: single-hop between  $E67\_Birth$  and  $E21\_Person$  and multi-hop between  $E67\_Birth$  and the second  $E53\_Place$ .

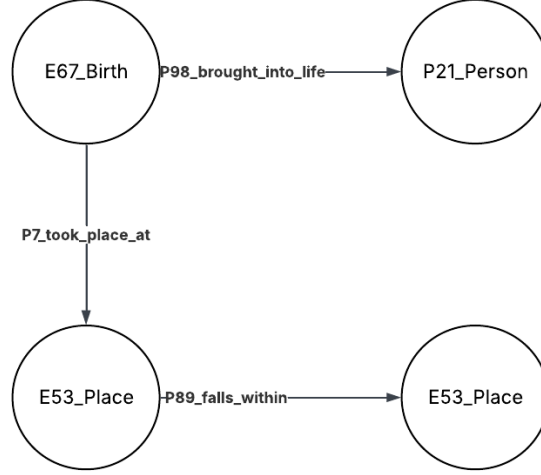


Figure B.1: Nodes and Edges of the Provided Schema Relations.

The observable neighbourhood via reachability for  $E67\_Birth$ :

$$U_{E67\_Birth} = \{E67\_Birth, E21\_Person, E53\_Place\}$$

The observable neighbourhood via reachability for  $E53\_Place$ :

$$U_{E53\_Place} = \{E53\_Place\}$$

Therefore, a valid example of a multi-hop CQ template using the above would be: *In which broader place does  $\{E67\_Birth$  of  $E21\_Person\}$  take place?*. The terminal class,  $E53\_Place$ , is in the observable neighbourhood of  $E67\_Birth$  and the resulting CQ template is as such structurally admissible. An example of how that multi-hop CQ template materialises when instantiated with an entity (Oscar Wilde), in other words, how the entity information is contained within the VRTI KG can be seen in Figure B.2. This would give the following filled-in question and answer: *In which broader place does the birth of Oscar Wilde take place? Leinster.*

An invalid example excluded under the proposed directional design constraints of this approach would be: *Which births occurred in a particular  $\{Place\}$ ?*. While this is a reasonable question and could be determined via SPARQL, it is invalid as it would require the traversal upstream with respect to the schema. This notion of observable neighbourhood aligns with the exploratory design principle of the Tús Maith framework: questions should extend outward from an entry entity along semantically meaningful directions rather than traversing the schema arbitrarily.

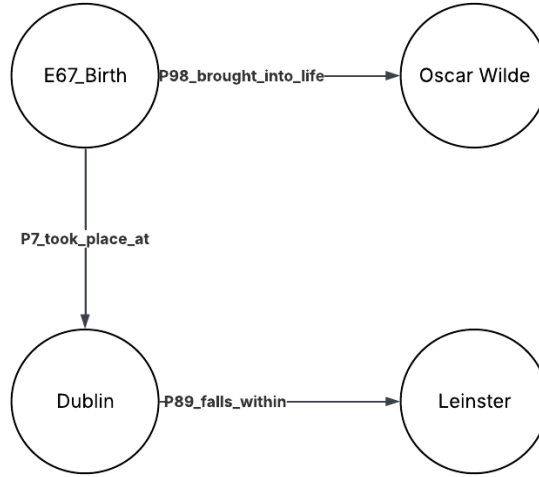


Figure B.2: Nodes and Edges instantiated by Oscar Wilde.

#### B.1.4 Stage 1 Augmented: Proposed System for Topology-Informed Multi-Hop CQ Template Generation

Rather than in Stage 1 of TMv1, providing the LLM solely with just the classes and properties of the KG, the LLM instead is provided with structural signals derived from the ontology schema. Specifically, schema-level statistics would be mined from the KG: degree profiles (number of in- and out-degrees), property cardinality patterns (functional (1-1), one-to-many (1-n), many-to-one (n-1), many-to-many (n-n)), and hierarchical depth. On the basis of these signals, building on [168, 169, 171], the proposed design introduces a schema traversal and selection component that operates over the directed class-property graph of the ontology. This component performs a bounded exploration of the schema graph, beginning from a designated entry class and iteratively extending class-property-class sequences (meta-paths) to a pre-defined hop depth. At each extension step, candidate continuations are evaluated against a set of structural admissibility criteria derived from the extracted signals. Specifically, continuations involving high-multiplicity relations (e.g., one-to-many, or many-to-many cardinality patterns) are treated as indicators of exploratory potential.

To preserve alignment with the entity-centric anchoring model of Tús Maith, the admissibility of each candidate meta-path is further constrained by the observable neighbourhood of the entry class, as defined by the Alexandroff topology induced by the schema preorder. Only those paths whose terminal classes lie within this neighbourhood are retained ensuring that the multi-hop CQ templates remain structurally reachable under the directed semantics of the ontology.

In traversing the schema graph in this manner, it is hypothesised that, in a similar manner to OnSET [28], it is possible to create SPARQL queries iteratively to align with the meta-paths for use later in Stage 2.

The retained meta-paths are then presented to the LLM as structured design constraints guiding the generation of natural language CQ templates that reflect both the multi-

relational structure of the ontology and the epistemic perspective as defined by the directionality of the entry class. As per TMv1, all generated CQ templates are still subject to domain expert review, potential edits, and ultimate selection before their subsequent instantiation. The criteria for CQ templates still holds in that they must be answerable by the KG, clear grammatically, and relevant as candidate exploratory starting points into the KG.

### **B.1.5 Stage 2 Augmented: Addressing Instantiation Challenges for Multi-Hop CQ Templates**

Multi-hop CQ templates reduce the likelihood of an entity chosen at random from the KG containing information in its immediate semantic neighbourhood that could instantiate the template, in other words, a sparsity of candidate entities challenge has been introduced. Therefore, instead of directly using popular entities to instantiate the multi-hop CQ templates, a cohort identification component is added as a precursor to the entity-ranking component. This now involves the requirement of the SPARQL-pair translation of the multi-hop CQ template i.e., the retrieval of entities whose information specifically matches the pattern outlined by the multi-hop CQ template. As hypothesised in the preceding section however, this SPARQL query would be already formed and so ready to be executed to return candidate entities to instantiate the CQ template. As per TMv1, the entity ranking component can then come into effect providing a list of entities to the domain expert that could be used to instantiate the multi-hop CQ template. In contrast to the original presentation of the entities to the domain expert which listed all entities for all CQ templates, the entities would be tied to a particular multi-hop CQ template. Again, as per TMv1, the LLM would still be passed the entity information and the multi-hop CQ template with the resulting filled-in questions and answers reviewed and curated by the domain expert. Furthermore, the validation component is hypothesised to still be a viable first-pass filter as the three checks still hold: is there a valid entity URI? does the entity name/names for the URI match the name slot/slots in the query? does the answer exist within the provided entity information?

In this way, while a SPARQL query is tied to a particular CQ template, it is just to surface candidate instantiation entities. At the time of the lay user following the multi-hop CuQA, the approved answer is presented to them and the general query about the entity or entities still fires thereby maintaining the entity-centric approach aligning with exploration vs. information retrieval. In other words, the decoupling of the narrow precision of the question from the broad recall of the exploration experience is still maintained.

### **B.1.6 Evaluation Methodologies for Complex CuQAs**

The proposed framework separates valid answerability (topological admissibility under the schema preorder) from usefulness (structural centrality and cardinality-based heuristics), establishing a evaluation framework in which formal structural admissibility (schema-theoretic validity) and epistemic usefulness (user-perceived exploratory value) can be independently realised and empirically assessed.

### **B.1.7 Future Work Conclusion**

The introduction of topology-informed constraints also introduces computational and conceptual complexity, and it remains an open question whether such formalisation yields sufficient practical benefit over simpler schema-traversal heuristics. Empirical comparison would therefore be required before adopting this approach within Tús Maith.