

The InterDev Framework: Harnessing Semantic Web Technologies to Enable FAIR Evidence in International Development

by

Matt Murtagh-White

Thesis

Presented to the University of Dublin, Trinity College,
in fulfilment of the requirements for the Degree of

Doctor of Philosophy

Supervised by Prof. Declan O'Sullivan

Co-supervised by Dr. P.J. Wall

University of Dublin, Trinity College

February 2026

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

I, the undersigned, 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).

Matt Murtagh-

White

28th February

2026

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

Matt Murtagh-

White

February 28th,

2026

Acknowledgments

I would like to thank my supervisors Prof. Declan O’Sullivan and Dr. P.J. Wall for their dedicated support throughout my PhD and the production of this thesis. I am grateful to Janet Choi at University College Cork for her work in developing the CRT-AI programme into an engaging and resourceful funding environment for its PhD researchers. My sincere thanks go to Cormac McKenna and Tom Carron of the ADAPT TechOps team, the truly unsung heroes of the research centre, without whose expertise and support the technical research into language models presented in this thesis would not have been possible. I would like to be in a position to thank Taighde Éireann beyond their mandatory requirements for us to do so. While this support enabled the completion of the work presented here, broader structural challenges surrounding doctoral research funding in Ireland, such as maternity leave and the recognition of research as professional work, remain an important area for continued improvement. The following is the aforementioned acknowledgement:

This thesis has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland under Grant No.18/CRT/6223.

Matt Murtagh-White

University of Dublin, Trinity College

February 2026

Abstract

International development research depends on the ability to discover, compare, and synthesise evidence from heterogeneous impact evaluations, systematic reviews, registries, and contextual datasets. In practice, however, this evidence remains fragmented across repositories, reporting formats, and levels of abstraction, making it difficult for researchers and practitioners to reuse findings efficiently or transparently. This thesis addresses that problem through the design, implementation, and evaluation of InterDev, an ontology-driven linked-data platform for managing and exploring international development evaluation evidence.

InterDev is built around an ERCT ontology-aligned knowledge graph that supports evidence browsing, faceted filtering, collection-based curation, structured submission, and export, without requiring users to possess semantic web expertise. The research follows a six-phase User-Centred Design process. Initial requirements were derived through a domain-grounding study based on a review of systematic reviews in international development, which identified recurring needs for advanced filtering and contextualisation, access to both granular and aggregated evidence representations, and support for cross-study comparison and harmonisation.

The thesis then reports an iterative evaluation programme across three design iterations. The first iteration established baseline usability and identified major interaction breakdowns in filtering, collection management, export, and manual submission. The second iteration introduced multi-criteria filtering, multi-collection workflows, improved export, and LLM-assisted submission. This substantially reduced manual contribution effort and improved completion rates relative to manual entry, while remaining compatible with human review and validation. The third iteration extended evaluation beyond usability to examine the epistemic effects of LLM-mediated exploration under different provenance conditions. Results show that

conversational mediation reduced perceived workload, but did not reliably improve task performance or efficiency once task difficulty was controlled. Surfaced provenance changed how users judged outputs and influenced confidence and verification behaviour, but did not by itself improve trust or confidence calibration.

To probe whether these interaction-level trade-offs were specific to international development or reflected broader properties of exploratory AI systems, the thesis also reports a complementary cross-domain evaluation in a digital-library setting. That study found comparable patterns: conversational interfaces reduced interactional effort and often increased user confidence, but did not eliminate the need for verification and careful synthesis. This cross-domain result is used as bounded supporting evidence for the broader relevance of the interaction design implications, without claiming generalisation of InterDev itself.

Together, the findings show that ontology-driven linked-data platforms are a viable basis for reusable evidence infrastructures in international development, and that LLM mediation can reduce interactional effort, particularly for contribution and exploratory access. However, the thesis also demonstrates that usability gains do not automatically translate into improved epistemic outcomes. The central design challenge is therefore not simply adding AI assistance, but integrating it in ways that preserve provenance, support verification, and enable calibrated, reliable evidence use.

Contents

Acknowledgments	vi
Abstract	vii
Contents	ix
List of Tables	xiv
List of Figures	1
1 Introduction	4
1.1 Motivation and Scope	4
1.2 Research Question	7
1.3 Research Objectives	9
1.3.1 Research Objective 1 (RO1)	9
1.3.2 Research Objective 2 (RO2)	9
1.3.3 Research Objective 3 (RO3)	10
1.3.4 Research Objective 4 (RO4)	11
1.4 Technical Approach	11
1.5 Evaluation Approach	12
1.6 Thesis Contributions	15
1.6.1 Major: InterDev	15
1.6.2 Minor: Operationalising and extending ERCT for evaluation evidence	18
1.6.3 Minor: Feasibility study of LLM-assisted submission	18
1.6.4 Minor: Empirical evidence on exploration, provenance cues, and user reasoning	19
1.6.5 Additional Publications that supported the research	21
1.7 Thesis Overview	23
1.7.1 Chapter 2: Background	23
1.7.2 Chapter 3: State of the Art	23
1.7.3 Chapter 4: Implementation	23
1.7.4 Chapter 5: Evaluation	23
1.7.5 Chapter 6: Conclusion	24
2 Background	25
2.1 Domain Context: FAIR, Linked Evidence and International Development	25
2.1.1 The evidence lifecycle in international development	25
2.1.2 Limits of FAIR in Development Evidence	28
2.1.3 Semantic Web Technology Stack in Relation to InterDev	30
2.2 ERCT In Context	31
3 State of the Art	35
3.1 State of the Art in Interfaces for Knowledge Graphs, LLMs, and Their Integration	35
3.1.1 Pillar 1: Evidence-synthesis workflows	38
3.1.2 Pillar 2: Provenance, traceability, and publication units	40
3.1.3 Pillar 3: Schema and ontology operationalisation for user-facing systems	41
3.1.4 Pillar 4: Knowledge graph interaction patterns for non-experts	43

3.1.5 Pillar 5: LLM and knowledge graph integration for extraction and retrieval	46
3.1.6 Pillar 6: Human–AI interaction for appropriate reliance	48
3.1.7 Summary: Design requirements and interaction principles	50
3.2 Domain grounding and requirements elicitation from systematic reviews	52
3.2.1 Understanding the Process: A Review of Systematic Reviews	53
3.2.2 Review Framework and Parameters	53
3.2.2.1 Sample	54
3.2.2.2 Phenomenon of Interest	55
3.2.2.3 Design	55
3.2.2.4 Evaluation	56
3.2.2.5 Research Type	56
3.2.3 Search Protocol and Strategy	56
3.2.3.1 Databases and Repositories Searched	57
3.2.3.2 Keyword Strategy	58
3.2.3.3 Selection Criteria	59
3.2.4 Findings from Review-of-Reviews: Understanding User Needs and Contexts	60
3.2.4.1 Filtering and Contextualisation of Data	61
3.2.4.2 Data Collection Practices	62
3.2.4.3 Focus on Intervention Impact	63
3.2.4.4 Contextual Challenges	63
3.2.5 Identification of Initial User Requirements	64
3.2.5.1 User Requirement 1: Advanced Filtering and Contextualisation	65
3.2.5.2 User Requirement 2: Access to Granular and Aggregated Data	66
3.2.5.3 User Requirement 3: Tools for Cross-Study Comparisons and Harmonisation	66
4 Implementation	70
4.1 Design Approach and User-Centred Process	70
4.2 System Architecture and Technology Stack	72
4.2.1 Client-Server Architecture	74
4.2.2 Knowledge Graph Backend	75
4.2.3 RDF Tooling and Validation	76
4.3 Data Sources and Knowledge Graph Construction	77
4.3.1 Source Datasets	77
4.3.2 Semantic Uplift and Ontology Alignment	79
4.3.3 Knowledge Graph Integration and Storage	79
4.4 Core Interaction Abstractions	80
4.4.1 Global Interface Structure	80
4.4.2 Evidence View	80
4.4.3 Collection View	81
4.4.4 Submission View	82
4.5 Iteration-Specific Implementation Extensions	82
5 Evaluation	84
5.1 Evaluation methodology for InterDev framework	86
5.2 Experimental Setup	87
5.2.1 Participant Setup	87
5.2.2 Participant Recruitment	88
5.3 Evaluation Instruments	88
5.3.1 Post Study System Usability Questionnaire	89
5.3.2 Evaluation Metrics	89
5.3.2.1 Quantitative Measures	89
5.3.2.2 Qualitative Measures	90
5.3.2.3 Supplementary measures introduced by iteration	90

5.3.3 Data Analysis.....	91
5.3.3.1 Quantitative analysis	92
5.3.3.2 Qualitative analysis	92
5.4 Usability Evaluation Iteration 1	94
5.4.1 System Under Evaluation	94
5.4.2 Experimental Design	94
5.4.3 Quantitative Results.....	97
5.4.4 Qualitative Results.....	99
5.4.5 Limitations	100
5.4.6 Conclusions.....	101
5.5 Usability Evaluation Iteration 2	102
5.5.1 Updated User Requirements.....	102
5.5.2 Evaluating LLM Assisted Knowledge Graph Generation from Text.....	104
5.5.2.1 Context: Automated KG Construction and Document Parsing	105
5.5.2.2 Approach	106
5.5.2.3 Results.....	109
5.5.2.4 Conclusions	110
5.5.3 Changes to System Design.....	111
5.5.3.1 Filtering and Navigation Enhancements (Requirement 1)	111
5.5.3.2 Data Download Process (Requirement 2)	111
5.5.3.3 Collection Management Features (Requirement 3)	112
5.5.3.4 Impact Evaluation Submission Workflow Updates (Requirement 4)	112
5.5.3.5 LLM-Assisted Data Submission Pipeline.....	113
5.5.4 Ontology changes as a result of iteration 1.....	115
5.5.5 Experimental Design	116
5.5.5.1 Experimental Design Modifications	117
5.5.5.2 Participant Setup and Recruitment.....	117
5.5.5.3 Task List and Evaluation Instruments	118
5.5.6 Results	119
5.5.6.1 PSSUQ	119
5.5.6.2 Task Timing	120
5.5.6.3 LLM Assisted Submission	121
5.5.6.4 User Trust and Perceived Accuracy	122
5.5.7 Limitations	123
5.5.8 Iteration 2: Conclusions	123
5.6 Usability Evaluation Iteration 3.....	124
5.6.1 Identified Requirements.....	126
5.6.2 Changes to System Design.....	128
5.6.2.1 Evidence View extensions introduced in Iteration 3	128
5.6.2.2 Multi-collection workflow as the unit of work.	129
5.6.2.3 Collection-bounded LLM mediation and experimental condition	130
5.6.2.4 Provenance-cue visibility manipulation.....	131
5.6.3 Ontology Changes as a result of iteration 2	131
5.6.4 User Study	133
5.6.4.1 Participants	133
5.6.4.2 Study Procedure	133
5.6.5 Measures.....	137
5.6.6 Results and Analysis.....	138
5.6.7 Quantitative Results.....	139
5.6.7.1 Task Performance and Efficiency	139
5.6.7.2 Workload (NASA-TLX)	140
5.6.7.3 Confidence and Calibration.....	142
5.6.7.4 Trust in Automation.....	143
5.6.7.5 PSSUQ	143
5.6.7.6 Summary of Quantitative Patterns Findings	144
5.6.8 Qualitative Results.....	145

5.6.8.1 Summary of Qualitative Patterns Findings	147
5.6.9 Limitations.....	147
5.6.10 Iteration 3: Conclusions	148
5.6.11 Ontology changes as a result of iteration 3.....	152
5.7 FOOPS Review of Ontology.....	153
5.8 Cross-Domain Evaluation of Interaction Modalities	156
5.8.1 Introduction	157
5.8.2 System Design	159
5.8.2.1 GutGet: Grounded Retrieval-Augmented Recommendations Over Project Gutenberg	159
5.8.2.2 C0: Keyword Search.....	163
5.8.2.3 C1: Semantic Search.....	164
5.8.2.4 C2: Chat Search.....	165
5.8.2.5 C3: Conversation-Chat Search	166
5.8.3 User Study	166
5.8.3.1 Participants.....	167
5.8.3.2 Conditions	167
5.8.3.3 Study Procedure	168
5.8.3.4 Data Processing.....	170
5.8.3.5 Measures.....	171
5.8.3.6 Analysis	171
5.8.4 Results	172
5.8.4.1 Quantitative Results and Analysis	172
5.8.4.2 Qualitative Results.....	177
5.8.5 Discussion	180
5.8.5.1 Chat as interactional offloading, with synthesis as the remaining bottleneck	180
5.8.5.2 Balancing Exploratory Power and Cognitive Load in Hybrid Visual–Conversational Interfaces	181
5.8.5.3 Limitations.....	182
5.8.6 Cross Domain Evaluation: Conclusion.....	183
6 Conclusion	185
6.1 Summary of findings in relation to the research question.....	186
6.1.1 What was demonstrated.....	186
6.1.2 What was partially demonstrated	186
6.1.3 What remains unresolved.....	187
6.1.4 Overall summary.....	187
6.2 Findings in relation to the research objectives.....	187
6.2.1 Research Objective 1: State of the art and derived requirements.....	187
6.2.2 Research Objective 2: InterDev design, implementation, and formative evaluation.....	188
6.2.3 Research Objective 3: LLM-mediated interaction and evaluation across domains.....	189
6.2.4 Research Objective 4: ERCT refinement and ontology quality assessment.....	190
6.3 Design implications for linked-data evidence platforms with LLM mediation	191
6.4 Future Work	193
6.4.1 Longitudinal deployment in real evidence-synthesis settings.....	193
6.4.2 Claim-level verification and calibrated reasoning support	194
6.4.3 Semantic robustness in LLM-assisted submission.....	194
6.4.4 Scalable support for synthesis over bounded evidence collections	194
6.4.5 Extending the assessment protocol to AI-supported evidence synthesis.....	195
6.5 Closing Statement.....	195
Bibliography.....	197
Appendix	221
A. InterDev Study Iteration 3 Protocol.....	221

B.	<i>Project Gutenberg Cross Domain Study Protocol.....</i>	230
C.	<i>End-to-End Prompt Example for Project Gutenberg Retriever.....</i>	243

List of Tables

Table 1: Canonical model commitments in the ERCT Ontology	33
Table 2: Literature reviewed	38
Table 3: Repositories and Databases used for Review of Systematic Reviews	58
Table 4: Selection Criteria for Review of Systematic Reviews	60
Table 5: Initial User Requirements identified	69
Table 6: Data and Implementation details across InterDev iterations	72
Table 7: Stable platform infrastructure and iteration specific experimental interaction layers	73
Table 8: Overview of InterDev Evaluation Iterations	85
Table 9: Evaluation measures and instruments used throughout evaluation iterations and the cross-domain study	91
Table 10: Task list and associated requirements	96
Table 11: Thematic analysis summary for the first iteration of InterDev user testing. ...	100
Table 12: Comparison of BLEU and Macro-F1 scores for RDF subgraph generation from evaluation reports	109
Table 13: Updated task list and associated requirements for second evaluation.	118
Table 14: Metrics for LLM-Assisted Submission and Manual Workflows Submission.	122
Table 15: Tasks completed by participants during the evaluation	136
Table 16: Participant prompts used for the post-evaluation interview.	137
Table 17: Summary of qualitative themes, key findings, reference counts, and analytic scope.	145
Table 18: FOOPS assessment results for ERCT V5, comparing the baseline V5 run (overall score 0.9000) with the final remediated run (overall score 0.9333), reported as Pass or Failure per check.	155
Table 19: Evaluation tasks and their target behaviours. Each participant completed two distinct tasks per interface block	170
Table 20: Summary of qualitative themes identified from interviews and open-ended responses, including their key findings, reference counts, and analytical scope	177

List of Figures

Figure 1: InterDev Evaluation Framework	13
Figure 2: Initial ERCT ontology structure, showing the core trial-centred model and its main ontology properties linking interventions, outcomes, methodology, timespan, publication information, researchers, and provenance.	32
Figure 3: InterDev Methodological Process - Understanding User Needs and Contexts, Defining User Requirements.....	53
Figure 4: InterDev Methodological Process - Prototyping and Design	71
Figure 5: Overview of Framework	74
Figure 6: Overview of Implementation Methodology	77
Figure 7: Evidence view of InterDev Version 1	81
Figure 8: InterDev Methodological Process – Usability Testing, Implementation and Continuous Evaluation.....	86
Figure 9: Boxplot of time spent to complete each task in the usability evaluation.	97
Figure 10: Bar chart of the number of interventions required for each task.....	98
Figure 11: Boxplot of PSSUQ scores, based on a Likert 7-point scale where lower values indicate a higher satisfaction. System Usefulness (SysUse), information Quality (InfoQual) and Interface Quality (IntQual) and Overall are aggregated metrics based on questions 1-8, 9-15, 16-18 and 1-19 respectively.	99
Figure 12: Prompt structure for extracting RDF triples from evaluation reports.	108
Figure 13: Iteration 2 Evidence View with multi-criteria filtering and collection actions.	111
Figure 14: Iteration 2 Collection View, enabling multi-collection management.	112
Figure 15: Iteration 2 Submission View with LLM-assisted pre-population from an uploaded evaluation report.....	113
Figure 16: Iteration 1 ERCT ontology refinements, showing the retained core relations alongside the new canonical facet properties and annotation properties used to stabilise browsing, filtering, and interface terminology.	115
Figure 17: Distribution of PSSUQ Item Scores and Subscale Averages.....	119
Figure 18: Task Durations Distribution	121
Figure 19: The evidence view of the third iteration of InterDev. The evidence view now contains additional categories for faceted browsing (left) and text search (top).	129
Figure 20: Collection View of the Third Iteration of InterDev. Collections are user defined subgraphs of evidence from the ERCT knowledge graph.	130
Figure 21: Inside a collection on the third iteration of InterDev. The " LLM-mediated exploration (no provenance cues)" condition includes an LLM that answers to user questions by retrieving data from the underlying user-defined subgraph.....	131

Figure 22: Iteration 2 ERCT ontology extensions, showing the submission and extraction provenance properties, typed submission fields, source-document linkage, and SKOS-backed facet properties added to support LLM-assisted ingestion and controlled reuse.	132
Figure 23: Task performance by task and condition. C0 indicates the baseline condition, C1 LLM-mediated exploration (no provenance cues) and C2 LLM-mediated exploration with provenance cues.	139
Figure 24: Task Completion Time by Task and Condition. Time is log transformed to better fit the linear model. C0 indicates the baseline condition, C1 LLM-mediated exploration (no provenance cues) and C2 LLM-mediated exploration with provenance cues.	140
Figure 25: Overall Aggregated NASA-TLX workload score. C0 indicates the baseline condition, C1 LLM-mediated exploration (no provenance cues) and C2 LLM-mediated exploration with provenance cues (* $p < .05$, ** $p < .01$, *** $p < .001$).	140
Figure 26: NASA-TLX workload subscales, C0 indicates the baseline condition, C1 LLM-mediated exploration (no provenance cues) and C2 LLM-mediated exploration with provenance cues (* $p < .05$, ** $p < .01$, *** $p < .001$).	141
Figure 27: Self-Reported Confidence and Brier confidence calibration measures. C0 indicates the baseline condition, C1 LLM-mediated exploration (no provenance cues) and C2 LLM-mediated exploration with provenance cues (* $p < .05$, ** $p < .01$, *** $p < .001$).	142
Figure 28: Jian et al Trust in Automation measure. C0 indicates the baseline condition, C1 LLM-mediated exploration (no provenance cues) and C2 LLM-mediated exploration with provenance cues.	143
Figure 29: Distribution of PSSUQ item scores and aggregated metrics (SysUse, InfoQual, IntQual, Overall). Lower scores are better.	144
Figure 30: Distribution of qualitative code frequencies across participants, organised by overarching theme. Darker cells indicate higher code counts, and marginal totals summarise the number of references per code and per participant.	146
Figure 31: Iteration 3 ERCT ontology, showing the finding-level evidence model, finding provenance, evidence signalling properties, lexical alignment properties, and export/effect-reuse properties added to support verification and synthesis.	152
Figure 32: The GutGet backend pipeline. User prompts are parsed into structured intents constrained by a Gutenberg ontology (a), used to retrieve and filter book metadata from Gutendex with an LLM relevance gate (b), and presented through grounded explanations and an interactive graph with chat (c).	159
Figure 33: Overview of the original Project Gutenberg interfaces: (A) Homepage highlighting keyword search bar, (B) Example of a search results page (searching "travel", sorted by popularity), (C) Main categories listing covering literature, history, arts, and sciences, (D) Advanced search interface allowing multi-field queries (author, title, subject, language, etc.).	163
Figure 34: Overview of the original Project Gutenberg website interfaces of the book detail: introduction, download links, categories and other detailed data.	164
Figure 35: Overview of our Semantic Search and an example query results for "books about travel in 19th century": (A) no results returned with keyword search before searching, (B) relevant results retrieved with contextual search.	164

Figure 36: Overview of the Conversational Search interface: (A) User initiates a query ("give me books about travel"), (B) System responds with book recommendations and hyperlinks, (C) User follows up with another request ("recommend one of your book recommendations, explain why"), (D) System provides a detailed explanation supporting the recommendation.....	165
Figure 37: Overview of our Conversational-Graph Search Interface (C3) with an example query “books about travel”: (A) initial query with no results yet, (B) contextual results retrieved and visualized in a relation graph, (C) user selection of one book and showing the metadata of the book, (D) system highlighting relationships between selected books.	166
Figure 38: Mean outcomes by interface condition. The first two panels show rubric scores for synthesis/comparison and interpret/verify tasks (0 – 7), the right panel reports lookup/filter accuracy. Error bars represent standard errors (* $p < .05$, ** $p < .01$, *** $p < .001$).....	173
Figure 39: Mean rubric synthesis task scores (0–7). Error bars represent standard errors (* $p < .05$, ** $p < .01$, *** $p < .001$)	174
Figure 40: Mean completion time across task types. Error bars represent standard errors (* $p < .05$, ** $p < .01$, *** $p < .001$)	175
Figure 41: Self-reported confidence scores by task type. Error bars represent standard errors (* $p < .05$, ** $p < .01$, *** $p < .001$).....	175
Figure 42: Brier score by task type. Error bars represent standard errors (* $p < .05$, ** $p < .01$, *** $p < .001$).....	176

Chapter 1

Introduction

1.1 Motivation and Scope

International development is generally understood as a set of activities aimed at improving the welfare of people at risk of poverty and addressing inequalities within and across societies (Concern Worldwide, 2022). Yet many international development projects, particularly those involving information systems and technology, fail either totally or partially (Heeks, 2002). Explanations for this include a persistent gap between “Northern Designs” and “Southern Realities” in ICT interventions (Heeks, 2002), while Fryer argues that knowledge about what works in development often accumulates gradually through repeated successes and failures (Fryer Jr, 2017).

In parallel, international development has increasingly adopted data-driven evaluation approaches, commonly labelled as Randomised Controlled Trials (RCTs) or impact evaluations¹ (Banerjee et al., 2016). These approaches implement a policy intervention on a randomised sample from a representative population and estimate treatment effects on welfare-related outcomes such as consumption, educational attainment, and employment. Because of their controlled and randomised design, impact evaluations are often valued for identifying causal effects rather than merely recording correlations (Reiley & List, 2007). More broadly, they embed monitoring, data gathering, and evaluation within development projects in order to identify which policies and levers are most effective (Sachs, 2012).

Many individual evaluations have generated important insights. For example, evaluations embedded in cash transfer programmes such as Mexico’s PROGRESA showed that childhood exposure to transfers can improve educational attainment, labour market outcomes, and household investment in productive assets (de Walque et al., 2017; Kugler

¹ This thesis will use the term impact evaluations to describe these studies.

& Rojas, 2018; Parker & Vogl, 2018; Schultz, 2004). Likewise, research on school-level deworming programmes found positive spillover effects not only on health but also on educational outcomes for school-age children (Ahuja et al., 2018; Miguel & Kremer, 2004).

These studies have substantially expanded the evidence base for data-driven evaluation, and such methods are now common in development economics (Banerjee et al., 2016). They have also generated a large and growing volume of context-specific evidence: together, the 3ie Development Evidence Portal and the American Economic Association RCT Registry list more than 16,000 evaluations (American Economic Association, 2021; International Initiative for Impact Evaluation, 2022).

However, growth in the number of evaluations has not been matched by consistent data availability or replicability (Gertler et al., 2018; Kopper et al., 2020; Miguel, 2021). This matters because cross-study synthesis, including replication, reanalysis, and meta-analysis, depends on access to well-documented data and comparable metadata across settings (Bandiera et al., 2016; Kopper et al., 2020; Meager, 2019). In response, research organisations have adopted data publication policies and FAIR-aligned guidance, where FAIR refers to making research outputs more Findable, Accessible, Interoperable, and Reusable (Abdul Latif Jameel Poverty Action Lab, 2020; Bill and Melinda Gates Foundation, 2021; Wilkinson et al., 2016). Yet existing evidence repositories largely remain per-evaluation silos with inconsistent metadata structures and limited support for linking across datasets, constraining interoperability and increasing the cost of synthesis (American Economic Association, 2021; Harvard College, 2023; International Initiative for Impact Evaluation 2021). Because contributing structured metadata is typically optional and depends on author self-report, data entry and harmonisation remain labour-intensive and are often incomplete. In other words, improving reuse in practice requires not only better metadata standards and linked representations, but also interactional solutions that make structured evidence easier to contribute, inspect, and verify.

Knowledge graphs have been proposed as a response because they provide a shared identifier-based representation with explicit relations and provenance, improving interoperability and traceability (Chaves-Fraga et al., 2022; Kamdar & Musen, 2021; Manghi, 2024; Sun et al., 2022). The remaining bottleneck is practical: researchers must be

able to populate, inspect, and reuse graph data efficiently, motivating interaction designs and assisted workflows that reduce manual structuring effort while maintaining user oversight (Dadzie et al., 2011; Ye et al., 2024).

Within this research gap, **this thesis proposes InterDev**, an ontology-driven linked-data platform that (i) represents evaluation metadata and outcomes as an ontology-aligned knowledge graph, (ii) provides interactive interfaces for browsing, filtering, and curating evidence without requiring semantic web expertise, and (iii) supports low-burden contribution through human-in-the-loop workflows, while enabling export to familiar tabular formats for downstream analysis.

Recent work suggests that combining conversational interaction with graph-based exploration can reduce the interaction cost of navigating large corpora, but it also raises questions about verification when outputs are mediated by Large Language Models (LLMs) (P. Jiang et al., 2023; Steyvers et al., 2025; Ye et al., 2024). Motivated by this tension, the thesis evaluates InterDev both as a usable evidence platform and as an interaction approach for reliable, verifiable exploration, including a cross-domain probe.

These conditions motivate two coupled design problems. First, evaluation evidence requires a shared, identifier-based representation that can link impact evaluations, interventions, outcomes, datasets, and reports with explicit provenance so that evidence can be discovered and reused across repositories. Second, that representation must be operationalised through interaction: researchers must be able to populate, filter, inspect, and curate evidence without semantic-web expertise, and where LLMs are used must be able to verify outputs rather than accept them on faith. The novelty of this thesis lies in combining an ontology-driven evidence representation, user-centred interaction design for non-expert evidence work, and bounded LLM mediation evaluated not only for usability, but also for verification and confidence calibration. This thesis addresses both problems through InterDev, and evaluates whether the resulting interaction approach improves practical reuse while supporting reliable reasoning.

1.2 Research Question

The research question posed is:

To what extent can an ontology-driven linked-data platform, augmented with LLM-mediated interaction, improve the interoperability and practical reuse of international development impact evaluations while supporting reliable and verifiable user reasoning?

The terms used in the research question can be explained as follows:

Ontology-driven linked-data platform: A system that represents evaluation evidence using a shared conceptual model (an ontology) and identifier-based structured data so that entities (e.g., impact evaluations, interventions, outcomes, contexts) can be linked consistently across sources and queried as an integrated corpus. A Linked Data Platform can further refer to a set of best practices and technologies that enable structured data to be shared and interconnected on the web, following a standardised approach (Bizer et al., 2011; Van Hooland & Verborgh, 2014; Zinke-Wehlmann et al., 2021). In this thesis, the ontology provides schema constraints that govern ingestion, linking, querying, and user-facing browsing and filtering using the World Wide Web Consortium (W3C) standard Web Ontology Language (OWL) (McGuinness & Van Harmelen, 2004; Staab & Studer, 2013). This is supported by additional standards in this technical stack such as the Resource Description Framework (RDF) (Brickley et al., 2014; Klyne, 2004) and the SPARQL query language used to fetch and edit RDF data (DuCharme, 2013; Harris, 2013). The specific ontology used in this case is ERCT (Evaluation, Intervention, Context, Trial), an ontology describing the methods, concepts and outcomes of impact evaluations. ERCT was initially developed in the author's previous MSc. research and iterated further as part of this thesis (Murtagh-White, 2021).

Interoperability and practical reuse: Guided by the FAIR principles, a set of guidelines aimed at improving the management and sharing of data, ensuring that data are Findable, Accessible, Interoperable, and Reusable (Wilkinson et al., 2016). This thesis targets evidence that is findable and accessible, and that supports cross-study integration and reuse through comparable, machine-actionable metadata.

International Development: Commonly understood as a set of efforts aimed at enhancing the well-being of individuals vulnerable to poverty, with the primary goal of addressing and reducing inequalities both within and between societies. (Concern Worldwide, 2022)

Impact Evaluations: In international development, impact evaluations are systematic assessments of programs, projects, or policies to measure their effectiveness, relevance, and impact, usually through the use of a control group created via staggered program rollouts (Glennester & Takavarasha, 2014). These evaluations help determine whether development goals are being met and provide insights for improving future interventions.

LLM-mediated Interaction: The use of large language models (LLMs) to mediate user interaction with the platform by generating or proposing structured content and natural-language responses under human oversight (Bommasani et al., 2021; Gilardi et al., 2023). In this thesis, LLM-mediated interaction has two roles. **LLM-assisted submission** supports submission by extracting and proposing ERCT-aligned metadata from unstructured sources for user review and correction prior to ingestion. **LLM-mediated exploration** supports discovery and synthesis by enabling natural-language querying over bounded evidence collections, with outputs grounded in retrieved knowledge-graph records and, where enabled, accompanied by provenance cues.

Reliable and verifiable user reasoning: the extent to which the platform enables users to form and justify evidence-based judgments over evaluation records such that (i) claims produced during exploration are grounded in the retrieved evidence set rather than unconstrained generation, (ii) users can inspect and, where needed, challenge those claims by following an explicit provenance path from statements to the underlying impact evaluations (and associated metadata), and (iii) users' subjective confidence in their conclusions is appropriately aligned with objective task performance.

1.3 Research Objectives

The research objectives (ROs) derived from the research question of this thesis (section 1.2) are outlined here.

1.3.1 Research Objective 1 (RO1)

RO1: Conduct a state-of-the-art review of methods and systems for representing, constraining, and interacting with evidence using knowledge graphs, provenance models, and LLM-mediated interaction workflows, and complement this with domain-grounded requirements elicitation from systematic review practice in international development, in order to derive design requirements and interaction principles for evidence synthesis.

This objective synthesises two inputs that jointly inform InterDev’s design. First, it surveys the state of the art across six pillars: evidence-synthesis workflows, provenance and publication units, schema and ontology operationalisation, non-expert KG interaction patterns, LLM–KG integration for extraction and retrieval, and human–AI interaction for appropriate reliance. Second, it grounds these insights in international development practice through a review-of-systematic-reviews (which is referred to as review-of-reviews henceforth), treating systematic reviews as traces of real workflow demands. The outputs are (i) a set of initial, domain-grounded system requirements for interoperable and provenance-aware evidence representations and contribution pipelines, and (ii) interaction principles for discovery, comparison, and verification, which together serve as the baseline for RO2–RO4. The fulfilment of RO1 is detailed in Chapter 3.

1.3.2 Research Objective 2 (RO2)

RO2: Design, implement, and formatively evaluate InterDev as an ERCT-aligned linked-data platform that enables non-semantic web researchers to browse, filter, curate, and export evaluation evidence from a knowledge graph.

This objective develops a researcher-facing interface for international development evaluation data structured using the Evaluation and Randomised Controlled Trial (ERCT) ontology, originally introduced during the author’s MSc research (Murtagh-White, 2021). Building on prior requirements work from the construction of the ERCT ontology derived from expert input and evidence-synthesis practice (Matt Murtagh-White et al., 2023; Murtagh-White, 2021) and then further developed from the review-of-reviews described in RO1, initial UI and API prototypes were implemented against ERCT-aligned test data and iteratively refined through user-centred studies, summarised in Section 1.5 and detailed in Chapter 5. The resulting interface supports structured browsing and faceted filtering, collection-based curation, and export for downstream analysis, enabling researchers to interrogate relationships across studies without requiring semantic web expertise. Design and refinement were guided by the needs identified in RO1.

1.3.3 Research Objective 3 (RO3)

RO3: Evaluate how LLM-mediated interaction changes evidence-exploration and submission behaviour in linked-data settings, including its effects on workload, confidence and confidence calibration, verification behaviour, and trust, under controlled interface and provenance conditions, and assess the portability of these effects in a cross-domain probe.

This objective evaluates how introducing LLM-mediated interaction changes user behaviour and judgment in linked-data evidence systems. It includes controlled comparisons of LLM-assisted submission against manual contribution, and controlled comparisons of LLM-mediated exploration interfaces that vary mediation and provenance visibility, assessing efficiency, workload, confidence calibration, verification practices on evidence synthesis-oriented tasks (i.e. in which users must combine, compare, and reconcile evidence from multiple evaluations or findings in order to form a coherent higher-level interpretation, answer, or judgment). A complementary cross-domain study in a digital library setting serves as a bounded external-validity probe, testing whether modality-driven effects observed in InterDev reflect broader interaction properties rather than domain-specific factors.

1.3.4 Research Objective 4 (RO4)

RO4: Further refine the Evaluation and Randomised Controlled Trial (ERCT) ontology in conjunction with additional data needs and the user evaluation.

This objective focuses on refining the ERCT ontology (Murtagh-White, 2021) to better meet the evolving data requirements identified throughout the research. Based on insights gathered from the user evaluations and the challenges encountered during the transformation of unstructured text data into RDF format, the ontology was expanded and adjusted. This refinement process ensures that the ontology more accurately captures the necessary entities, relationships, and properties relevant to international development evaluations. The aim was to align the ontology with both the data ingestion needs and the practical applications of users, ensuring it remains a robust and flexible tool for managing evaluation data within a linked-data framework and conforms to established quality standards such as FOOPS! (Garijo et al., 2021), an extension of OOP!s (Poveda-Villalón et al., 2012) that incorporates FAIR-oriented checks.

Taken together, RO1 establishes the design space and requirements; RO2 realises these as an implemented, researcher-facing platform; RO3 tests the behavioural and epistemic effects of LLM mediation under controlled interface and provenance conditions; and RO4 ensures that the underlying ontology remains fit for these workflows as empirical findings accumulate. The next sections describe the technical and evaluation approach used to deliver and test these objectives.

1.4 Technical Approach

InterDev was developed to address two linked challenges in the development-evidence ecosystem: the fragmentation of evaluation artefacts across repositories, and the difficulty non-specialist researchers face in using linked-data systems effectively. To address these challenges, the platform adopts a decentralised linked-data architecture in which evaluation artefacts are represented as an ERCT-aligned knowledge graph. This provides a shared

schema for integrating heterogeneous evidence and supports cross-study comparison through identifier-based linking, structured querying, and provenance-aware retrieval.

On top of this graph layer, InterDev provides a researcher-facing interface for browsing, filtering, curating, and exporting evaluation evidence without requiring semantic web expertise. The system therefore combines ontology-driven representation with interaction mechanisms designed around the evidence-synthesis tasks and user requirements derived in Chapter 3, and motivated by the evidence-lifecycle challenges outlined in Chapter 2.

Later iterations incorporate two LLM-mediated workflows. First, LLM-assisted submission proposes ERCT-aligned metadata from unstructured source materials for user review before ingestion. Second, LLM-mediated exploration supports natural-language querying, comparison, and inspection over bounded evidence collections, with responses grounded in retrieved graph records and, where appropriate, accompanied by provenance cues.

Alongside platform development, the ERCT ontology was refined iteratively in response to emerging data requirements, ingestion challenges, and user-study findings. Together, these components form the technical basis for the implementation described in Chapter 4 and the evaluations reported in Chapter 5.

1.5 Evaluation Approach

The evaluation of InterDev assesses not only usability but also whether the system better supports exploration and comparison in synthesis-oriented tasks and appropriately calibrated trust when users interact with evaluation evidence. While it incorporates standard usability study measures such as efficiency and error rates, this thesis evaluates how the aforementioned LLM-mediated workflows influence users' verification behaviour, workload, confidence, and trust calibration during evidence exploration and synthesis tasks.

This evaluation programme is guided by the six principles of User-Centred Design (UCD) in ISO 9241-210:2010, including understanding users and contexts, continuous user involvement, and iterative refinement based on feedback (DIS, 2009). These principles are operationalised through the six-phase process shown in Figure 1, which serves as the organising framework for the evaluation activities described throughout the thesis. Later chapters revisit individual phases as they become relevant and cross-reference this section for the full framework description.

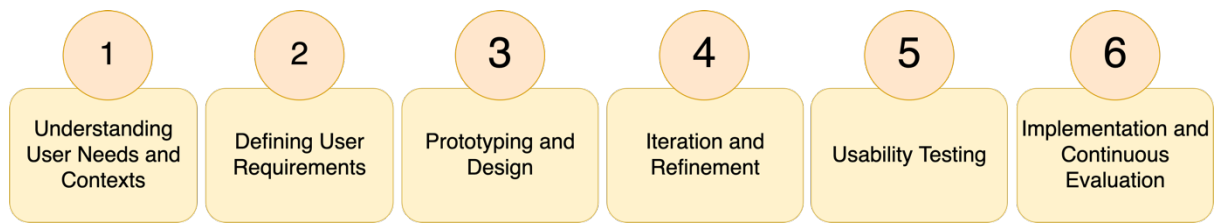


Figure 1: InterDev Evaluation Framework

1. Understanding User Needs and Contexts

Evaluation begins by characterising target users, their tasks, and the settings in which evaluation evidence is sought, compared, and reused. Methods at this phase include user and process research to ground design decisions in realistic workflows and constraints (DIS, 2009; Norman, 1988). The objective is to establish a baseline understanding of where evidence discovery and reuse breaks down in practice and what forms of support are plausible for domain researchers (Kujala, 2003). This phase is fulfilled in Chapter 3 with a review-of-reviews.

2. Defining User Requirements

Findings from Phase 1 are translated into explicit requirements that constrain both system functionality and evaluation design. Requirements specify the tasks the system must support (e.g., discovery, comparison, curation, export, and contribution) and provide criteria for assessing whether later iterations address identified breakdowns. This step reduces design drift by tying new features to observed needs rather than speculative additions (Preece et al., 2015). This phase is fulfilled in Chapter 3, with a mapping of initial user needs based on the outcome of the review-of-reviews.

3. Prototyping and Iterative Design

Requirements are operationalised through prototypes that externalise interaction concepts and enable early testing. Prototyping reduces the cost of change by allowing key assumptions about navigation, filtering, and curation to be evaluated before full implementation (Saffer, 2010). Prototypes are incrementally refined as feedback is gathered, ensuring that design decisions can be tested with representative users rather than inferred solely by designers (Nielsen, 1994).

4. Iteration and Refinement

InterDev is refined through successive iterations in which empirical findings from user studies directly motivate redesign. This phase links observed interaction breakdowns and user feedback to concrete changes in the interface, workflows, and supporting automation (e.g., filtering and collection management, export affordances, and contribution burden). Iteration is treated as an evidence-driven cycle: issues identified in one round motivate modifications, and subsequent studies assess whether those modifications reduce the same failures under comparable task conditions.

5. Usability Testing and Study-based Evaluation

Formal user studies are conducted on each iteration using task-based evaluation. Participants complete defined tasks reflecting key workflows (e.g., locating relevant evaluations, assembling comparison sets, exporting data, or contributing metadata), while behavioural and subjective measures are recorded. A small set of core instruments is retained across iterations to support comparability, including task completion time and the Post-Study System Usability Questionnaire (PSSUQ) (J. R. Lewis, 1992, 2002), supplemented by open-ended questionnaire responses analysed using a consistent thematic analysis procedure (Nowell et al., 2017).

As the research focus evolves across iterations, additional measures are introduced to address iteration-specific questions. In early formative testing, concurrent verbalisation (think-aloud) and assistance/intervention counts are used to localise interface (Boren & Ramey, 2000; Ericsson & Simon, 1980, 1998). In later iterations examining LLM-mediated interaction, the evaluation incorporates controlled experimental conditions and additional measures such as perceived accuracy and trust in the LLM-assisted submission workflow, and later for the LLM-mediated exploration tasks used NASA–TLX workload, task-level confidence ratings, calibration metrics (e.g., Brier-style scoring), and rubric-based scoring of synthesis quality under controlled provenance conditions (Hart & Staveland, 1988; J. Li et al., 2025; Ma et al., 2024a). The exact instrument set used in each iteration is specified in Chapter 5 (Section 5.3 and Table 9).

6. Implementation and Continuous Evaluation

Following study-based testing, improvements are implemented and evaluated in use over time, with feedback mechanisms used to detect emergent issues and evolving requirements as the evidence base and workflows change (ISO/TC 159/SC 4 Ergonomics of Human-System Interaction (Subcommittee), 1998).

In parallel with the user-facing evaluation programme, the ERCT ontology is assessed against the FAIR principles using FOOPS (Garijo et al., 2021). The FOOPS evaluation is used to identify modelling pitfalls and guide refinement, with attention to clarity of conceptual modelling, logical consistency, reuse of established vocabularies, and FAIR-aligned interoperability. This ontology-level quality assurance supports the ingestion, linking, and querying operations that InterDev depends on.

1.6 Thesis Contributions

This thesis makes four contributions. The central contribution is the design, implementation, and iterative evaluation of the InterDev framework: an end-to-end linked-data platform intended to improve how international development impact evaluation evidence is structured, accessed, curated, and reused. Three additional contributions support this primary outcome: (i) an updated and operationalised version of the ERCT ontology embedded within InterDev; (ii) a technical study demonstrating the feasibility and limitations of LLM-assisted submission for generating ontology-aligned knowledge graph content from unstructured evaluation reports; and (iii) empirical findings on how LLM-mediated exploration and provenance cues shape usability, workload, confidence, and verification behaviour in knowledge-graph-based evidence use, including a cross-domain probe in digital libraries.

1.6.1 Major: InterDev

The major contribution of this thesis is InterDev, an open-source framework comprising an ERCT-aligned knowledge graph, a user interface for evidence browsing and curation, and two forms of LLM-mediated interaction: LLM-assisted submission and LLM-mediated exploration. The requirements for InterDev are informed by previous work based on the ERCT ontology (Murtagh-White, 2021) which highlighted a need for more structured ways of accessing and amalgamating evaluation data within International Development and the

review-of-reviews detailed in Chapter 3. Additional rounds of usability testing allowed for user feedback to be incorporated into the platform, with changes and improvements made based on each feedback round.

InterDev was evaluated across three iterations. The first iteration established baseline usability and identified core interaction breakdowns, particularly in filtering, collection management, export affordances, and manual submission. The second iteration addressed these issues through multi-criteria filtering, multi-collection curation, improved export formats, and an LLM-assisted submission workflow aimed at reducing the cost and friction of structured contribution. The third iteration extended evaluation beyond task completion and satisfaction, examining how LLM-mediated interaction reshapes evidence exploration on synthesis-oriented tasks by measuring cognitive workload, confidence and its calibration, and verification behaviour under controlled provenance conditions.

Outputs from this contribution include:

- Matt Murtagh-White, Declan O’Sullivan, P.J. Wall, From Submission to Synthesis: Evaluating LLM-Mediated Knowledge-Graph Workflows for Development Evidence Curation. Under review for Semantic Web Journal, 2026.

This paper reports Iteration 2 and Iteration 3 of this thesis (Sections 5.5 and 5.6), examining how LLM assistance shapes interaction with knowledge-graph-based evidence systems across two workflows: ontology-aligned data submission and exploratory evidence use over curated collections. The studies combine controlled experimental designs to assess effects on usability, workload, confidence, trust, and judgment quality. Results show that LLM-assisted workflows can substantially reduce interactional burden and lower barriers to contribution, while also altering how users evaluate, verify, and rely on system outputs. In particular, the findings highlight a recurring trade-off between efficiency gains and epistemic engagement: automation improves usability and perceived ease of use, but does not automatically improve reasoning quality, and transparency mechanisms such as provenance cues can shift confidence and verification behaviour without straightforwardly increasing trust. Overall, the work characterises the benefits and limits of LLM-mediated interaction for knowledge graph curation and evidence synthesis, and motivates design strategies that balance assistance with user oversight.

- Matt Murtagh-White, P.J. Wall, Declan O’Sullivan, Evaluating InterDev: A FAIR Platform for International Development Data, In Proceedings of the 9th International Workshop on Visualization and Interaction for Ontologies and Linked Data co-located with the 19th International Semantic Web Conference, VOILA@ISWC 2024

This paper explains the structure, design and purpose of InterDev and details the initial user evaluation of the platform. This research comprises of the entirety of the Iteration 1 evaluation as described in Section 5.4.

- Matt Murtagh-White, P.J. Wall, Declan O’Sullivan, The ERCT Ontology: Impact Evaluations, Ontology and Policy, Presented at the Development Studies Association of Ireland (DSAI) Conference 2022

This presentation of in-progress research of the ERCT was presented at the ICT4D track at the DSAI’s 2022 conference in Limerick to other development professionals and policymakers in this area.

- Matt Murtagh-White, InterDev: A FAIR Platform for International Development Data [Software], 2024, GitHub repository, <https://github.com/mmurtaghw/InterDevResearch>

The InterDev public repository contains both the frontend and backend components required to create a local instance of InterDev or develop the platform further.

InterDev’s intended impact is to reduce the practical cost of making evaluation evidence reusable across studies by shifting work from manual cleaning and harmonisation toward structured curation and analysis. For international development researchers and practitioners, this supports more reliable cross-study comparison and synthesis by enabling interoperable browsing, export, and contribution workflows aligned with FAIR principles. For the research community, InterDev provides an operational, open-source reference architecture demonstrating how ontology-driven knowledge graphs and LLM-assisted, user-centred interaction can be combined to support verifiable evidence exploration over fragmented corpora.

1.6.2 Minor: Operationalising and extending ERCT for evaluation evidence

InterDev incorporates an updated and system-integrated version of ERCT, an ontology developed by the author to represent key entities and relations in impact evaluation evidence, including interventions, contexts, methods, and outcomes. This thesis contributes by (i) further refining ERCT to better support evidence discovery and cross-impact evaluation comparison and (ii) demonstrating how an ontology can function not only as a descriptive artefact, but as an operational backbone for user workflows such as filtering, collection building, export, and structured submission.

An impact of the thesis research in this area can be demonstrated through the following outputs:

- Matt Murtagh-White, P.J. Wall, Declan O’Sullivan, Learning from the Evidence: Impact Evaluations, Ontology and Policy, In 2023 IEEE 17th International Conference on Semantic Computing (ICSC) (pp. 104-106). IEEE.

This paper describes the construction of the ERCT ontology and presents an evaluation. The ontology can be found on the thesis GitHub repository².

1.6.3 Minor: Feasibility study of LLM-assisted submission

A supporting contribution is a technical study evaluating the feasibility of using LLMs to extract structured, ontology-aligned representations from unstructured evaluation reports. This work investigated whether general-purpose LLMs can generate ERCT-aligned RDF subgraphs from report text, and quantifies trade-offs between lexical similarity, semantic accuracy, and syntactic validity of generated RDF. The study motivated the design of InterDev’s LLM-assisted submission workflow by showing that model outputs can potentially reduce manual entry effort while still requiring validation and human review to ensure standards compliance and data quality.

The impact of this evaluation can be seen in the following publications:

² <https://github.com/mmurtaghw/InterDevResearch>

- Matt Murtagh-White, P.J. Wall, Declan O’Sullivan, PRICER: Leveraging Few-Shot Learning with Fine-Tuned Large Language Models for Unstructured Economic Data, In Second International Workshop on Semantic Technologies and Deep Learning Models for Scientific, Technical and Legal Data (SemTech4STLD) co-located with the Extended Semantic Web Conference 2024 (ESWC 2024)

This paper demonstrates the process of fine-tuning a Mistral 7B open source LLM to take in raw HTML text data describing prices from an online retailer and categorise it as RDF within the Classification of Individual Consumption by Purpose (COICOP), a United Nations international standard for categorising data for the Consumer Price Index (CPI).

- Matt Murtagh-White, *PRICER: Leveraging Few-Shot Learning with Fine-Tuned Large Language Models for Unstructured Economic Data* [Dataset], Zenodo repository, <https://doi.org/10.5281/zenodo.10993684>. (2024)

This dataset supports the paper presented at the Second Workshop on Semantic Technologies and Deep Learning Models for Scientific, Technical, and Legal Data at ESWC 2024, describing the taxonomy used for economic data.

1.6.4 Minor: Empirical evidence on exploration, provenance cues, and user reasoning

This thesis contributes empirical findings about how LLM-mediated interaction changes user experience and judgment in knowledge-graph-based evidence systems. Iteration 2 provides evidence that LLM-assisted submission can increase task completion rates and reduce completion time relative to manual entry, while also shifting perceived accuracy and trust in the submission workflow. Iteration 3 extends this analysis to LM-mediated exploration evaluated through exploration, comparison, and synthesis-oriented tasks, showing how chat-based mediation affects perceived workload, confidence, and confidence calibration, and testing whether surfacing provenance cues changes verification practices and reliance. A complementary digital-library study using a shared RAG backend with keyword, semantic, conversational, and conversational-graph interfaces provided a bounded external-validity probe to further compose this contribution.

The impact of this contribution can be seen in the following publications:

- Matt Murtagh-White, Nicole Leon, Yunkai Xu, Jovan Jeromala, and Hiba Al-Nabhani, *From Queries to Conversations: How Interface Modality Shapes Exploratory Search Behavior*. Under review for ACM UIST 2026 (39th Annual ACM User Interface Software and Technology), Detroit, USA, Nov 2-5, 2026.

This work compares keyword, semantic, conversational, and conversational-graph interfaces for exploratory search over Project Gutenberg using a shared LLM-based RAG backend. A controlled user study shows that conversational interaction improves synthesis quality and reduces perceived workload, but also increases self-reported confidence without corresponding accuracy gains, indicating risks of premature closure. Adding a graph view supports cross-item comparison but increases cognitive effort, highlighting trade-offs between exploratory power, usability, and confidence calibration.

- Matt Murtagh-White, Yunkai Xu, Nicole León, and Frank Ritter, *Athena: A Conversational Book Discovery System with LLM-Powered Retrieval-Augmented Generation (RAG) and Interactive Graph Visualization*, In The 38th Annual ACM Symposium on User Interface Software and Technology (UIST Adjunct '25), Busan, Republic of Korea, 2025. ACM. <https://doi.org/10.1145/3746058.3758400>

This paper presents Athena, a proof-of-concept conversational system for exploratory book discovery. Athena integrates LLM-powered retrieval-augmented generation with interactive graph-based visualisation to support open-ended search and sensemaking in digital library contexts, using the Project Gutenberg knowledge graph vis-à-vis the Gutendex API layer. The system demonstrates how conversational and visual interfaces can reduce cognitive burden, improve transparency, and scaffold exploratory discovery across fragmented datasets. While focused on digital libraries, the design principles illustrate the wider applicability of LLM-mediated, graph-interaction approaches to another domain. This work was produced as part of a six month EU funded research fellowship at Pennsylvania State University.

1.6.5 Additional Publications that supported the research

In addition to the contributions above, the following publications arose as part of the development of this thesis as described in the brief introduction of each paper:

- Jovan Jeromela, Matt Murtagh-White, Alok Debnath, Jasmina Gajcin, Dipto Barman, and Owen Conlan, “As Long as It Does What I Want, I’d Be Happy to Trust It”: Exploring User Perspectives on a Scrutable and Proactive Assistant for Time Management, under review for the 34th ACM Conference on User Modeling, Adaptation and Personalization (UMAP 2026).

This work examines how users evaluate trust, control, and transparency in proactive time-management assistance through SIPA-TM, a scrutable assistant that proposes calendar actions while exposing its reasoning and user-model factors. A mixed-methods user study shows how explanation granularity, perceived agency, and privacy expectations shape acceptance of proactive scheduling, and it yields design implications for scrutability that support delegation without undermining user control, motivating iteration 3’s focus on user behaviour and judgement in response to LLM-based interaction systems.

- Jason Lucas, Matt Murtagh-White, Adaku Uchendu, Ali Al-Lawati, Michiharu Yamashita, Dominik Macko, Ivan Srba, Robert Moro, and Dongwon Lee, BLUFF: Benchmarking the Detection of False and Synthetic Content across 58 Low-Resource Languages, under review for *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD ’26)*, Jeju, Korea, 2026. ACM.

This work introduces BLUFF, a large-scale multilingual benchmark for detecting false and synthetic content across 78 languages and multiple authorship conditions, and shows that current detection models degrade sharply in long-tail languages, exposing limitations in cross-lingual robustness and safety pipelines. In addition to these empirical contributions, the work developed substantial expertise in LLM generation pipelines, which directly informed and strengthened the implementation of this thesis.

- Jason Lucas, Matt Murtagh-White, Ali Al-Lawati, Adaku Uchendu, and Dongwon Lee, DIA-GUARD: AI Safety and Harmful Content Robustness Across 50 English, 64th Annual Meeting of the Association for Computational Linguistics (ACL '26), San Diego, USA, 2026. ACM.

This work introduces DIA-GUARD, a large-scale benchmark and evaluation framework for analysing the robustness of AI safety systems and harmful content detectors across 50 English dialects, showing that performance degrades substantially when inputs deviate from Standard American English, particularly for smaller models. It further proposes an LLM-based and rule-based dialect transformation pipeline and shows that dialectal variation acts as a natural adversarial perturbation, exposing systematic weaknesses and uneven protection in current AI safety and moderation pipelines. This work also developed substantial expertise in LLM generation pipelines, which directly informed and strengthened the implementation of this thesis.

- Matt Murtagh-White, P.J. Wall, and Declan O'Sullivan, How Low Can We Go? Quantization Effects on LLM SPARQL Generation, In The 22nd Extended Semantic Web Conference (ESWC '25 Poster), Portorož, Slovenia, 2025. Springer.

This work supported the development of an efficient RAG pipeline for InterDev by studying how LLMs translate natural language into SPARQL queries for graph-based data retrieval, and by evaluating the impact of quantization on performance in conversational querying of linked data. The results demonstrate clear trade-offs between efficiency and accuracy. The paper was nominated for the Best Poster Award at ESWC 2025.

- Matt Murtagh-White, Jeffrey Sardina, and Declan O'Sullivan, SPARQLStack: Optimizing SPARQL Query Generation with Tiered LLMs, In The 23rd Extended Semantic Web Conference (ESWC '26), Dubrovnik, Croatia, 2026. Springer. (under review)

This paper builds on the InterDev RAG pipeline by proposing a tiered LLM approach to natural language-to-SPARQL translation, balancing efficiency and accuracy through model stacking. This work extends the technical foundation for conversational interaction with knowledge graphs, ensuring that linked-data exploration can scale effectively across

domains.

1.7 Thesis Overview

The remainder of the thesis is structured as follows:

1.7.1 Chapter 2: Background

Chapter 2 establishes the domain and representational context for the thesis. It characterises the evidence lifecycle in international development, explains why FAIR is unevenly realised in practice for development evidence, and introduces the ERCT ontology as a deliberately minimal schema for linking evaluation artefacts with persistent identifiers and provenance.

1.7.2 Chapter 3: State of the Art

Chapter 3 contributes towards RO1 by presenting a survey of state-of-the-art work across six pillars relevant to InterDev: evidence-synthesis workflows, provenance and publication units, schema/ontology operationalisation, non-expert knowledge graph interaction patterns, LLM-KG integration for extraction and retrieval, and human-AI interaction for appropriate reliance. It then grounds requirements in international development synthesis practice via a review-of-reviews, translating observed workflow demands into initial system requirements that constrain Chapters 4–5.

1.7.3 Chapter 4: Implementation

This chapter describes the technical components common to all iterations of the InterDev framework in detail.

1.7.4 Chapter 5: Evaluation

This chapter describes the evaluation undertaken for the InterDev framework, an iterative evaluation that took place across three iterations. It details the structural and methodological aspects of the evaluation as well as an analysis of the results and evolving requirements throughout each iteration. It contributes a further cross-domain study that further investigates interaction modalities introduced in iteration 3.

1.7.5 Chapter 6: Conclusion

This chapter presents the key findings of the research described in this thesis. It discusses the extent to which the research question was answered and considers how much the research objectives have been achieved. Potential further avenues of research are discussed and outlined.

Chapter 2

Background

Chapter 1 motivated the need for an evidence platform that improves interoperability and practical reuse of development evaluations while remaining usable to researchers and practitioners. Chapter 2 provides the background required to support this thesis’s approach: it first characterises the evidence lifecycle and the points at which reuse fails, then examines the limits of applying FAIR principles in development evidence ecosystems, and finally introduces ERCT as the minimal ontology used to represent and link evaluation artefacts in InterDev.

2.1 Domain Context: FAIR, Linked Evidence and International Development

2.1.1 The evidence lifecycle in international development

The production and reuse of evidence in international development can be understood as a lifecycle running from the design of an impact evaluation to its eventual synthesis in systematic reviews and meta-analyses. Although this resembles the research cycle in other empirical domains, development practice introduces distinctive challenges: fragmented infrastructures, heterogeneous artefacts, and uneven adherence to transparency norms. The literature describes this both as a stepwise cycle and as part of a broader evidence ecosystem (Browne, 2024; Stewart, 2021).

Registration. The lifecycle begins with the specification of study design, increasingly captured through prospective repositories such as the AEA RCT registry, a centralised platform maintained by the American Economic Association for documenting randomised evaluations in economics and related fields, or RIDIE, an initiative of 3ie that focuses on documenting development interventions. Registration provides a public record of intended interventions, outcomes, and analysis plans, and is intended to mitigate selective reporting. While registration has become more common, coverage across sectors and geographies is uneven, and many entries contain minimal information. Broader debates on impact evaluation also highlight tensions between design choices, external validity, ethics, and the institutional incentives that shape what gets registered (Akram-Lodhi, 2020; Barrett &

Carter, 2020; Bédécarrats et al., 2019). These inconsistencies motivate InterDev’s emphasis on a shared representation that can preserve links between registered intent, later outputs, and contextual metadata.

Data collection. Field implementation generates artefacts such as survey instruments, enumerator manuals, raw microdata, and monitoring dashboards. Their quality and documentation vary substantially, with direct implications for future reuse. Multi-sector and cluster-randomised evaluations further raise issues around unit of analysis, spillovers, and contamination, increasing the metadata burden required for comparison across settings. These challenges place high demands on metadata if findings are to be compared across settings (De Mel et al., 2009; Menon et al., 2020; Quisumbing et al., 2020). This motivates InterDev’s focus on structured, cross-study metadata that can carry design and context information forward into later stages of reuse and synthesis.

Analysis. The analysis stage transforms raw data into cleaned datasets, code, and statistical results. Transparent alignment between design and analysis is central to credibility, regardless of whether identification relies on randomisation or observational (Elbers & Gunning, 2014). In practice, reproducibility remains a concern. Audits of published international development evaluations find that only a minority are computationally reproducible using authors’ own data and code (Wood et al., 2018). The meaning of “replication” itself has been contested, with recent contributions distinguishing replication from re-analysis and documenting the social as well as technical dynamics of these exercises (Bédécarrats et al., 2021; Ozier, 2021). This motivates InterDev’s emphasis on explicit links between results, underlying datasets, and analytic provenance so that downstream users can inspect how findings were produced.

Reporting. Findings are disseminated through academic papers, policy briefs, and donor reports. Donor-mandated reporting often results in long narrative documents that contain valuable contextual information but are poorly structured for discovery or integration. Reporting incentives can also shape how findings are classified and framed, leaving divergences from registered plans, code, and data insufficiently visible (Kabeer, 2020; Michaelowa & Michaelowa, 2011). Without systematic linkage back to registered plans, code, and data, divergences remain opaque. This motivates InterDev’s approach of treating reports not as terminal narrative outputs, but as evidence artefacts that must remain linkable to data, methods, and source records.

Repository deposit. Some outputs are preserved in repositories such as data archives, journal replication packages, or curated evidence databases. Deposit practices, however, are inconsistent. Articles are often archived without underlying datasets, and datasets appear without persistent identifiers linking them to the interventions and outcomes they document. The rapid growth of impact evaluations since the mid-2000s has widened this gap between the quantity of studies and the reliability of the supporting artefacts (Cameron et al., 2016). Enforcement of open data and code policies has also been weak, with many studies failing to provide the materials required by journals (Wood et al., 2018). This motivates InterDev’s use of persistent identifiers and ontology-aligned links to reduce the disconnect between archived outputs and the evaluations they describe.

Reuse and synthesis. The final stage of the lifecycle involves the aggregation of studies in systematic reviews, meta-analyses, and evidence gap maps. It is here that weaknesses in upstream stages become most apparent. Reviews in international development often confront heterogeneity in design, missing or inaccessible datasets, and insufficient contextual metadata, limiting their ability to produce robust comparative insights (Mallett et al., 2012). Evidence gap maps provide one partial solution by mapping intervention–outcome pairs and highlighting areas of scarcity. The policy uptake of evidence has also been shown to depend on institutional capacities and incentives, not only on the supply of research outputs (Stewart, 2023). This motivates InterDev’s goal of making evaluation evidence discoverable, comparable, and contextually traceable across otherwise disconnected sources.

Across these stages, heterogeneity of outputs and uneven metadata emerge as defining problems. Artefacts differ in structure, terminology, and accessibility: registry records may be skeletal, datasets under-documented, reports unstructured, and repository links broken. The result is a persistent replication and availability gap: despite an exponential growth in published impact evaluations over the past two decades (Cameron et al., 2016), only a fraction are transparently documented, reproducible, and accessible for synthesis. This gap has led to the proliferation of transparency initiatives such as pre-analysis plans, mandatory data-sharing policies, and replication programs (Grant & Khatua, 2024), yet uptake remains partial.

2.1.2 Limits of FAIR in Development Evidence

The FAIR principles, findable, accessible, interoperable, reusable, were introduced in 2016 as a concise framework for improving the stewardship of digital research objects (Wilkinson et al., 2016). FAIR has since been widely adopted across disciplines and research infrastructures, providing a common vocabulary for data management and a reference point for funders and repositories (Boeckhout et al., 2018). When fully implemented, FAIR practices enable persistent identifiers, rich metadata, clear access conditions, standard vocabularies, and explicit licensing, together supporting reproducibility and reusability (Hughes et al., 2023). These ideals, however, have proven difficult to enact outside of well-resourced domains such as genomics and biomedicine. In social science and international development research, structural, technical, and cultural barriers complicate every pillar of FAIR.

Findability. FAIR requires that datasets be accompanied by sufficient metadata and persistent identifiers to support discovery. In practice, development research data often enter repositories with minimal or inconsistent documentation. Studies of FAIR implementation highlight the lack of incentives for researchers to invest time in metadata production, and the absence of common standards across repositories (Hughes et al., 2023). Fragmentation further undermines discovery: evaluations are scattered across evidence repositories, donor portals, and institutional archives, with no federated search. Multilingualism adds an additional layer of complexity in development contexts, as metadata produced in one language may not be discoverable to researchers in another (Dekker, 2020).

Accessibility. FAIR distinguishes between data that are openly accessible and data that are accessible under well-defined conditions. This distinction is especially relevant in international development, where microdata often contain sensitive information about individuals and communities. Although FAIR allows for authentication and authorization, the more detailed the metadata, the greater the risk of inadvertently revealing confidential information (Martorana et al., 2022). Researchers have called for tiered metadata and controlled access models, but implementation remains uneven. Accessibility is also constrained by infrastructure and geopolitics: Shanahan and Bezuidenhout (2022) demonstrate that physical location can determine whether “open” repositories are actually reachable, with connectivity and firewall restrictions limiting access from parts of the Global South.

Interoperability. The promise of FAIR lies partly in its capacity to enable data integration across disciplines. This pillar requires shared formats, vocabularies, and ontologies. In practice, social science data frequently fall short on interoperability indicators and disciplinary conventions continue to differ markedly (Batzdorfer et al., 2024). Even where ontologies or standards exist, aligning concepts across studies is complex: survey variables, outcome definitions, and classifications often diverge. Efforts to enforce uniform standards can meet disciplinary resistance, while bottom-up community approaches risk producing incompatible silos. In development evidence, heterogeneity in study design and reporting formats further compounds these problems. Without consistent semantic alignment and widespread adoption of community standards, interoperability remains largely aspirational.

Reusability. Reusability is the culmination of the other three principles, but it also introduces its own demands. To be reusable, data must be accompanied by detailed provenance, clear licensing, and sufficient context to allow secondary users to interpret them correctly. In development research, data are frequently shared without variable definitions, methodological documentation, or usage permissions. Even when files are deposited, they may be locked in proprietary formats or provided without annotation, rendering them unusable in practice (Batzdorfer et al., 2024; Wood et al., 2018). Moreover, as Boeckhout et al. (2018) observe, FAIR principles do not address normative dimensions such as ethics, equity, and benefit-sharing. These issues are particularly salient in international development, where secondary use of data can raise concerns about community rights, informed consent, and the fair distribution of benefits.

Taken together, these challenges explain why FAIR implementation is uneven in the evidence ecosystem described in Section 2.1.1. Findability is hampered by missing metadata and fragmented repositories. Accessibility is complicated by privacy concerns and infrastructural inequities. Interoperability is obstructed by heterogeneous standards and semantics. Reusability is undermined by incomplete documentation and unresolved ethical questions. For international development research, where outputs are heterogeneous, unstructured, and distributed across multiple actors, these barriers are amplified. The principles of FAIR offer a useful normative anchor, but “FAIR in practice” requires significant investments in infrastructure, standards, and incentives that are not yet in place. This thesis takes these limitations as a starting point, arguing that linked data approaches can help address some of the gaps by embedding persistent identifiers, provenance, and structured metadata into the evidence lifecycle.

These limitations imply that improvements to discovery and reuse in development evidence cannot rely on policy commitments to FAIR alone. They require machine-actionable representations that can carry identifiers, metadata, and provenance across heterogeneous artefacts, but also interfaces and workflows that make those representations usable for researchers who are not semantic web specialists. The remainder of this chapter therefore introduces the representational commitment taken in this thesis (the ERCT ontology), while Chapter 3 examines the interface and system patterns required to operationalise that commitment in practice.

2.1.3 Semantic Web Technology Stack in Relation to InterDev

To address the interoperability problems outlined above, this thesis adopts a semantic web technology stack in which RDF, OWL, and SPARQL play complementary roles. At the data level, the Resource Description Framework (RDF) provides a graph-based representation for expressing entities and relationships using persistent identifiers. This is important in the present context because international development evidence is fragmented across registries, repositories, reports, datasets, and contextual indicators that do not share a common structure. Representing these heterogeneous artefacts as RDF makes it possible to describe impact evaluations, interventions, outcomes, datasets, and contextual variables in a common machine-actionable form and to preserve explicit links between them (Brickley et al., 2014; Klyne, 2004).

On top of RDF, the Web Ontology Language (OWL) provides the shared conceptual model that gives those graph structures consistent meaning. In this thesis, that role is fulfilled by ERCT, which defines the principal classes and relations needed to represent evaluation evidence, such as evaluations, interventions, outcomes, datasets, contexts, and provenance-bearing links. OWL is therefore used here not primarily for abstract logical expressiveness in isolation, but as the schema layer that governs how heterogeneous source data are aligned, linked, and interpreted consistently across the platform (McGuinness & Van Harmelen, 2004; Staab & Studer, 2013). This is what allows InterDev to treat evidence from different repositories as part of a shared graph rather than as disconnected metadata silos, and it underpins later user-facing functions such as faceted browsing, collection-based curation, structured submission, and provenance-aware retrieval.

SPARQL completes this stack by providing the standard query and update language over RDF graphs. Within InterDev, SPARQL is used operationally rather than being exposed directly to end users: backend services translate interface actions such as filtering, browsing, collection management, and export into structured SPARQL queries and updates against the GraphDB triple store (DuCharme, 2013; Harris, 2013). This separation is important to the thesis because it allows researchers to benefit from structured graph retrieval without requiring semantic web expertise, while also preserving a stable and reproducible access pathway across evaluation iterations. In combination, RDF provides the representational structure, OWL provides the schema and semantics, and SPARQL provides the retrieval and manipulation mechanism. Together, these technologies form the linked-data foundation through which InterDev operates.

2.2 ERCT In Context

ERCT (Evaluation, Intervention, Context, Trial) is a deliberately compact domain ontology designed to provide a minimal, machine-actionable representation of impact evaluations that can be reused across evidence infrastructures. It builds on prior work by the author of this thesis, which developed an OWL ontology for impact evaluations in the social sciences (Murtagh-White, 2021), and reuses that project’s modular decomposition as the starting point for InterDev’s schema. ERCT is not intended as a technical solution for a single system. Rather, it is positioned as a pragmatic response to the socio-technical conditions described in Section 2.1: fragmented repositories, heterogeneous artefacts, weak linkage between outputs, and the difficulty researchers face when trying to discover, compare, and reuse evaluation evidence across settings. ERCT is therefore intended to support reliable linking between registry entries, datasets, code, results, and reports across the evidence lifecycle, rather than to model every variable or instrument in the social sciences. It prioritises a small number of stable classes and relations, while exposing linking points to richer domain vocabularies where needed.

Two complementary design commitments shape the ERCT model. First, minimal sufficiency: the schema must be able to answer a concise set of competency questions used by systematic reviewers and meta-analysts (see Table 1 and Figure 2 below); anything beyond what is required for those tasks is out of scope. Second, alignment-first: ERCT reuses mainstream vocabularies (for example, PROV-O for provenance, schema.org for

dataset descriptions) and exposes explicit mappings so that federation with other infrastructures is straightforward and semantic drift is minimised.

It is important to distinguish ERCT’s conceptual scope from the implementation scope of InterDev. Conceptually, ERCT is intended as a reusable ontology for representing core entities and relations in impact evaluation evidence across infrastructures. By contrast, the InterDev implementation operationalises only the subset of ERCT needed to support the platform’s requirements in this thesis, particularly browsing, filtering, curation, submission, and provenance-aware retrieval. In this sense, ERCT defines the wider representational model, while InterDev instantiates a task-driven subset of that model for a specific evidence platform context.

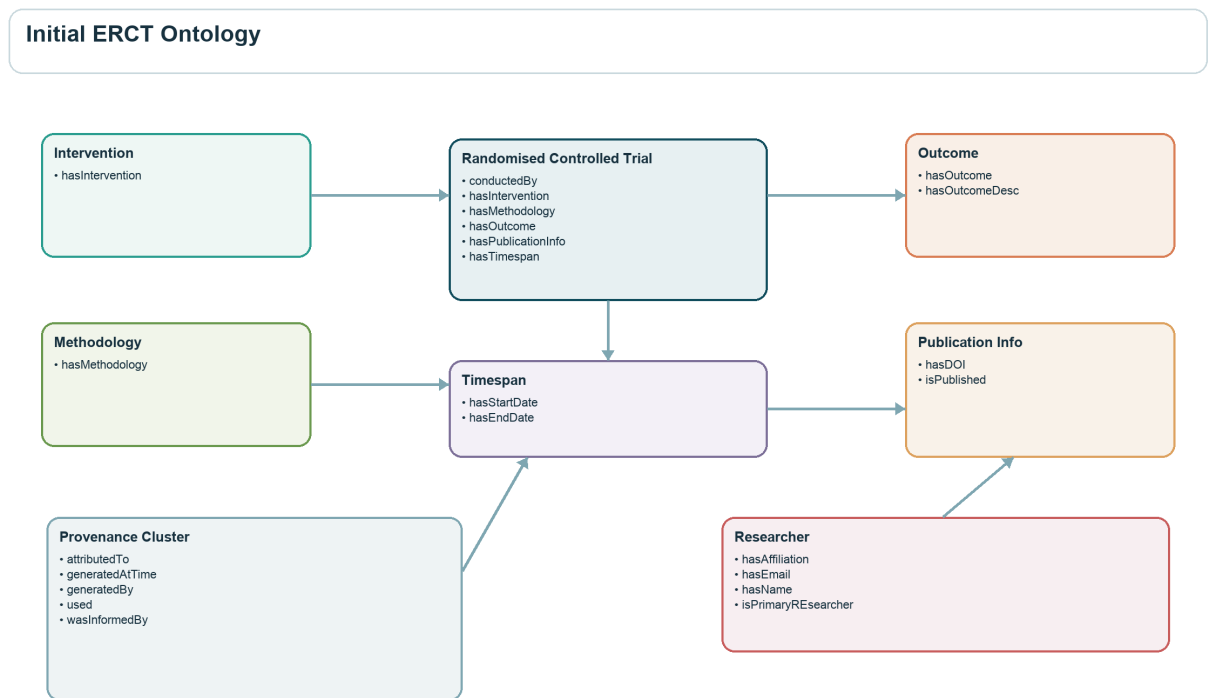


Figure 2: Initial ERCT ontology structure, showing the core trial-centred model and its main ontology properties linking interventions, outcomes, methodology, timespan, publication information, researchers, and provenance.

Core concept	Representative ERCT term and intent
Evaluation (impact evaluation record)	<code>erct:Evaluation</code> — central node linking interventions, outcomes, datasets and context
Intervention	<code>erct:Intervention</code> — a named policy/programme or treatment arm
Outcome	<code>erct:Outcome</code> — conceptual outcome (linked to operational measures)
Dataset	<code>erct:Dataset</code> — dataset artefact with distribution and licence metadata
Context	<code>erct:Context</code> — place/time/setting descriptors used for filters and covariates
Agent & Analysis	<code>erct:Agent</code> , <code>erct:Analysis</code> — actors and analytic artefacts
Provenance	PROV-O properties (<code>prov:wasDerivedFrom</code> , <code>prov:wasAttributedTo</code>) used for all assertions

Table 1: Canonical model commitments in the ERCT Ontology

Key relations include `erct:evaluates` (Evaluation $\rightarrow$ Intervention), `erct:hasOutcome` (Evaluation $\rightarrow$ Outcome), `erct:hasDataset` (Evaluation $\rightarrow$ Dataset), and `erct:conductedIn` (Evaluation $\rightarrow$ Context). The ontology also reserves slots for a small set of literal properties that are repeatedly required by meta-analysts: `startDate`, `endDate`, `sampleSize`, `randomisationUnit`, `primaryOutcome`, `license`, and an `extractionConfidence` score for automatically derived assertions.

The competency questions that shaped ERCT’s initial form were derived by the author during prior MSc research through (i) semi-structured interviews with three impact-evaluation experts and (ii) analysis of recurring workflow steps and query needs reported in systematic reviews and meta-analyses (Murtagh-White, 2021). These included identifying evaluations of a named intervention in a given country or time period, locating associated datasets and replication materials, finding evaluations that reported a specified outcome and effect estimate, and tracing a result back to its source artefact and extractor. Together, these questions required ERCT to provide persistent identifiers, explicit provenance links, and machine-tractable outcome descriptors without attempting to model variable-level nuance exhaustively. In this sense, ERCT was shaped not only by formal ontology design concerns, but by anticipated patterns of evidence use and reuse. The ontology is refined further in this thesis in response to identified user requirements.

ERCT therefore defines what InterDev must be able to represent and retrieve, but it does not determine how researchers will work with that representation. In practice, ERCT-

aligned data still need to be discovered, filtered, curated into working sets, and inspected with traceable links back to source artefacts, tasks that are fundamentally socio-technical as much as representational. This motivates the focus of Chapter 3: to review state-of-the-art approaches to knowledge graph interfaces, schema operationalisation, provenance surfacing, and LLM-mediated interaction, including LLM-assisted submission and LLM-mediated exploration, and to derive design implications for how an ERCT-backed platform can support evidence synthesis workflows without sacrificing auditability.

Chapter 3

State of the Art

Chapter 2 described the evidence lifecycle in international development and showed why FAIR-aligned reuse remains difficult in practice: key artefacts are fragmented across infrastructures, metadata quality is uneven, and linkages between registrations, data, code, and reported results are often missing or unstable. It also introduced the ERCT ontology as a deliberately minimal ontology intended to support persistent linking and retrieval across this lifecycle. Chapter 3 builds on these constraints by surveying the state of the art in interfaces for knowledge graphs (KGs), LLMs and their integration. The aim is not only to summarise recent technical and interactional advances, but to extract design implications that informed InterDev’s implementation and evaluation, specifically around interoperability, usability for non-experts, auditability via provenance, and appropriate reliance when LLM assistance is introduced. Section 3.1 covers the state of the art in interfaces for Knowledge Graph, LLMs and their integration and derived InterDev system design requirements based on the analysis of the state of the art. Section 3.2 presents a study based on a review-of-reviews in international development which derives domain-grounded user requirements for InterDev. In other words, this chapter operationalises ROI by mapping the relevant interface, representation, and human–AI interaction literature onto the concrete constraints identified in Chapters 1–2.

3.1 State of the Art in Interfaces for Knowledge Graphs, LLMs, and Their Integration

The literature reviewed here was selected to inform the specific technical and design requirements of InterDev (see Research Question and ROs in Chapter 1, Sections 1.2–1.3). Because InterDev is an applied platform that must (a) ingest and harmonise fragmented evaluation artefacts, (b) support researcher-facing exploration and curation, and (c) provide auditable provenance and trust signals when LLMs are involved, the review is organised around six pillars that align with these design needs. Pillars 1–3 address interoperability foundations for evidence-centric systems, covering evidence-synthesis workflows,

provenance, traceability and publication units for representing evidence artefacts, and schema and ontology operationalisation for constraining and governing user-facing data. Pillar 4 surveys interaction patterns that enable non-experts to explore and curate knowledge graphs, including visual exploration, faceted filtering, query builders, and natural language interfaces. Pillar 5 focuses on LLM and KG integration for extraction and retrieval, including retrieval-augmented generation and graph-guided reasoning architectures. Pillar 6 examines human–AI interaction for appropriate reliance, with particular emphasis on how interfaces shape trust calibration, verification behaviour, and the effective use of transparency cues.

Sources were identified through targeted searches in Google Scholar, ACM, IEEE Xplore, SpringerLink, and arXiv using keywords such as “knowledge graph interface”, “ontology visualisation”, “retrieval-augmented generation”, and “LLM explainability”. These searches returned a substantially larger candidate set than is reported here. Table 2 therefore lists the core sources used to structure the six pillars: papers selected because they are either (i) foundational/seminal for the representational and provenance mechanisms InterDev relies on, or (ii) directly evaluative of interaction patterns and reliance issues that constrain InterDev’s interface and study design. Older sources are retained where they define widely adopted primitives (e.g., named graphs, PROV-O) or establish canonical findings about graph readability and interaction.

Author (Year)	Paper Title	Category
Page et al. (2021)	<i>The PRISMA 2020 statement: an updated guideline for reporting systematic reviews</i>	(1) Evidence-synthesis workflows
Shamseer et al. (2015)	<i>PRISMA-P 2015: elaboration and explanation</i>	(1) Evidence-synthesis workflows
Rethlefsen et al. (2021)	<i>PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews</i>	(1) Evidence-synthesis workflows
Ouzzani et al. (2016)	<i>Rayyan—A web and mobile app for systematic reviews</i>	(1) Evidence-synthesis workflows
van de Schoot et al. (2021)	<i>ASReview: active learning for systematic review screening</i>	(1) Evidence-synthesis workflows
Tsafnat et al. (2014)	<i>Systematic review automation technologies</i>	(1) Evidence-synthesis workflows

Author (Year)	Paper Title	Category
Marshall & Wallace (2019)	<i>Toward systematic review automation: practical guidance on using machine learning tools in research synthesis</i>	(1) Evidence-synthesis workflows
Campbell et al. (2023)	<i>The “Big Picture” review family: mapping reviews, scoping reviews, and evidence gap maps (EGMs)</i>	(1) Evidence-synthesis workflows
Carroll et al. (2005)	<i>Named graphs, provenance and trust</i>	(2) Provenance / traceability
Groth, Gibson & Velterop (2010)	<i>The anatomy of a nanopublication</i>	(2) Provenance / publication units
Kuhn & Dumontier (2014)	<i>Trusty URIs: verifiable, immutable, and permanent digital artifacts for linked data</i>	(2) Provenance / publication units
Lebo et al. (2013)	<i>PROV-O: the PROV ontology</i>	(2) Provenance / traceability
Heath & Bizer (2011)	<i>Linked Data: Evolving the Web into a Global Data Space</i>	(2) Provenance / publication units
Staworko et al. (2015)	<i>Complexity and Expressiveness of ShEx for RDF</i>	(3) Schema/ontology operationalisation (
Poveda-Villalón et al. (2014)	<i>OOPS! (Ontology Pitfall Scanner): identifying common pitfalls in ontologies</i>	(3) Schema/ontology operationalisation
Garijo et al. (2021)	<i>FOOPS!: the FAIR Ontology Pitfall Scanner</i>	(3) Schema/ontology operationalisation
Dadzie & Rowe (2011)	<i>Approaches to visualising linked data: a survey</i>	(4) Interaction patterns
Kriglstein (2009)	<i>User requirements analysis on ontology visualization</i>	(4) Interaction patterns
Dudáš et al. (2018)	<i>Ontology visualization methods and tools: a survey</i>	(4) Interaction patterns
Katifori et al. (2007)	<i>Ontology visualization methods—a survey</i>	(4) Interaction patterns
Ghoniem, Fekete & Castagliola (2004)	<i>Readability of graphs: node-link vs. matrix</i>	(4) Interaction patterns
Liang et al. (2021)	<i>Querying knowledge graphs in natural language</i>	(4) Interaction patterns
Ferré (2016)	<i>Sparklis: An expressive query builder for SPARQL with natural language guidance</i>	(4) Interaction patterns
Arenas et al. (2016)	<i>Faceted search over RDF-based knowledge graphs</i>	(4) Interaction patterns
Francart et al. (2022)	<i>Sparnatural: a visual knowledge graph exploration tool</i>	(4) Interaction patterns

Author (Year)	Paper Title	Category
Dubey et al. (2019)	<i>LC-QuAD 2.0: A Large Dataset for Complex Question Answering over Wikidata and DBpedia</i>	(4) Interaction patterns
Verborgh et al. (2016)	<i>Triple Pattern Fragments: A low-cost knowledge graph interface for the Web</i>	(4) Interaction patterns
Yasunaga et al. (2021)	<i>QA-GNN: Reasoning with language models and knowledge graphs for question answering</i>	(5) LLM+KG integration
Pramanik et al. (2024)	<i>UNIQORN: unified QA over RDF knowledge graphs and natural language text</i>	(5) LLM+KG integration
Matsumoto et al. (2024)	<i>KRAGEN: A KG-enhanced RAG for biomedical problem solving</i>	(5) LLM+KG integration
Markowitz et al. (2024)	<i>Tree-of-Traversals: Zero-shot reasoning with knowledge graphs</i>	(5) LLM+KG integration
Reynolds & McDonell (2021)	<i>Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm</i>	(5) LLM interfaces supporting extraction/RAG
Nourani et al. (2019)	<i>Effects of meaningful and meaningless explanations on trust and perceived system accuracy</i>	(6) Appropriate reliance
Simhi et al. (2025)	<i>Trust Me, I'm Wrong: High-Certainty Hallucinations in LLMs</i>	(6) Appropriate reliance
Obrezkov et al. (2023)	<i>An Approach to Evaluate User Interfaces in a Scholarly Knowledge Communication Domain</i>	(6) Human–AI / evaluation for reliance

Table 2: Literature reviewed

3.1.1 Pillar 1: Evidence-synthesis workflows

As discussed in section 2.1, synthesis in applied research domains is organised around a structured but operationally demanding workflow that spans corpus construction, screening, data extraction and coding, and finally synthesis and reporting. Contemporary methodological guidance, most notably the PRISMA 2020 statement, formalises this process as a sequence of transparent and auditable steps intended to ensure reproducibility and methodological rigour (Page et al., 2021). Complementary extensions such as PRISMA-P, which focuses on protocol specification (Shamseer et al., 2015), and PRISMA-S, which standardises the reporting of search strategies (Rethlefsen et al., 2021), further

reinforce the centrality of workflow discipline, documentation, and traceability in modern evidence synthesis.

In practice, this workflow is increasingly supported by specialised tools, particularly at the screening stage. Systems such as Rayyan provide collaborative interfaces for managing inclusion and exclusion decisions, resolving reviewer conflicts, and coordinating large screening efforts (Ouzzani et al., 2016). More recently, machine-learning-assisted approaches such as ASReview have introduced active learning to prioritise likely relevant records for human assessment, demonstrating the effectiveness of human-in-the-loop paradigms for reducing workload without relinquishing expert control (Van De Schoot et al., 2021). Surveys of systematic review automation technologies and practical guidance for their adoption converge on a similar conclusion: computational methods are most useful when they augment specific, well-defined subtasks rather than attempting to replace the end-to-end review process (Marshall & Wallace, 2019; Tsafnat et al., 2014).

Despite these advances, the benefits of tool support remain unevenly distributed across the evidence-synthesis pipeline. While screening interfaces and prioritisation algorithms improve efficiency and consistency, data extraction and coding continue to rely heavily on manual processes and bespoke representations. Methodological standards treat extraction as a critical transformation step in which heterogeneous primary studies are mapped onto a shared analytical schema (Page et al., 2021; Shamseer et al., 2015), yet in practice the resulting datasets are typically stored in project-specific spreadsheets or databases. These representations are optimised for immediate analysis but are poorly suited for reuse, integration, or long-term curation across synthesis efforts. In other words, even when upstream workflow steps are well documented, the extracted evidence itself often remains difficult to re-query, recombine, or audit beyond the immediate project context.

The diversification of review types into scoping reviews, mapping reviews, and evidence gap maps further accentuates this limitation. Such approaches shift emphasis from single quantitative syntheses toward characterising the structure and coverage of the evidence landscape itself, for example by mapping interventions against outcomes or populations (Campbell et al., 2023). However, they remain constrained by the same underlying issue: intermediate products of evidence synthesis, such as screening decisions, extracted variables, coding schemes, and mapping structures are rarely treated as interoperable,

machine-actionable artefacts that can persist across studies or be reused as cumulative infrastructure.

Taken together, the literature reveals a persistent gap between well-specified workflows and the representations used to enact them. Standards such as PRISMA and its extensions define rigorous procedural expectations, and a growing ecosystem of tools supports individual stages of the pipeline, particularly screening. Yet the knowledge produced during evidence synthesis remains largely siloed within project-specific tools and formats. For platforms aiming to support cumulative, cross-study evidence building, this fragmentation poses a fundamental challenge. It motivates the need for infrastructures that prioritise structured, reusable, and auditable representations of extracted evidence and curated subsets, rather than attempting to automate the full review pipeline. The subsequent pillars address this representational bottleneck from the perspectives of provenance, schema operationalisation, and knowledge graph-based interaction.

3.1.2 Pillar 2: Provenance, traceability, and publication units

The limitations identified in current evidence-synthesis workflows are not only procedural but representational: while standards and tools increasingly structure how evidence is collected and processed, the intermediate and final artefacts of synthesis are rarely modelled in ways that support systematic reuse, verification, or recombination across projects. This gap has long been recognised in the Semantic Web community, where questions of provenance, trust, and publication granularity have been central to the design of data-centric infrastructures. Early work on named graphs introduced the idea that statements should be grouped into identifiable, addressable contexts, enabling assertions to be associated with sources, authorship, and trust metadata rather than existing as decontextualised triples (J. J. Carroll et al., 2005). This shift provides a conceptual foundation for treating evidence-synthesis outputs as addressable units rather than transient analysis tables. In a similar vein, the nanopublication model proposes packaging minimal scientific claims together with their provenance and publication information, thereby defining a fine-grained unit of scholarly communication that can be cited, inspected, and recombined across systems (Groth et al., 2010).

Subsequent work on verifiable identifiers further strengthens this notion of publication units by addressing reference stability and integrity. “Trusty URIs”, for example, embed

cryptographic hashes into identifiers, making digital artefacts verifiable and immutable once published (Kuhn & Dumontier, 2014). In evidence-synthesis settings, these mechanisms motivate the need for durable references to extracted findings, coded variables, and curated subsets of a knowledge base, even where full cryptographic immutability is not operationalised. Provenance models provide the formal glue that connects publication units to the processes and agents that produced them. The PROV ontology (commonly known as PROV-O) offers a general, machine-interpretable framework for describing entities, activities, and agents, and the relationships between them, such as derivation, attribution, and usage (Lebo et al., 2013). While originally motivated by data-intensive scientific workflows, this abstraction maps naturally onto evidence-synthesis pipelines, where screening decisions, extraction steps, transformations, and aggregations all constitute meaningful activities whose outputs should remain inspectable and auditable.

Taken together, this literature reframes data publication from releasing datasets to releasing structured, contextualised knowledge units with explicit links to sources and transformation steps. For evidence synthesis, this has two implications. First, intermediate artefacts, such as extracted outcome values, coding decisions, and curated collections, can be treated as first-class research objects rather than ephemeral by-products of analysis. Second, provenance and citation mechanisms provide an auditable basis for verification, but their effect on user trust and reliance is not guaranteed and must be evaluated at the interface level. This motivates InterDev’s treatment of curated collections as bounded evidence units and its emphasis on traceable links between interface outputs and underlying impact evaluations, forming a bridge from workflow concerns (Pillar 1) to the interaction and evaluation questions addressed in later chapters.

3.1.3 Pillar 3: Schema and ontology operationalisation for user-facing systems

While provenance models and publication units address how evidence artefacts can be contextualised, cited, and verified, they do not by themselves guarantee that such artefacts are structurally consistent, semantically coherent, or usable in practice. For knowledge graph-based systems that aim to support data entry, curation, and downstream querying, this places the burden on schemas and ontologies to serve not only as conceptual models but also as operational constraints on what counts as valid, meaningful data.

Constraint-based validation languages provide a formal mechanism for bridging this gap between conceptual modelling and executable system behaviour. Shape Expressions (ShEx), for example, define a schema language for RDF that allows graph structures to be validated against explicit shape constraints, with well-defined formal properties and complexity characteristics (Staworko et al., 2015). Such approaches shift ontologies from being purely descriptive artefacts to being enforceable specifications that can guide data entry, transformation, and quality control. In user-facing systems, this is particularly important: without enforceable constraints, interfaces may accept structurally inconsistent or semantically incomplete submissions, undermining both interoperability and trust in the resulting knowledge base. In practice, however, operational constraint enforcement can be realised through multiple layers, including ontology-aligned form models, controlled vocabularies for field values, and pre-ingestion validation gates that reject invalid serialisations before they reach the datastore.

Beyond structural validity, the long-term sustainability of ontology-driven systems depends on systematic quality assurance of the vocabularies themselves. The OOPS! framework identifies recurrent modelling pitfalls in ontologies and provides automated checks to detect them, highlighting issues such as missing domain and range declarations, inconsistent hierarchies, or inadequate annotation practices (Poveda-Villalón et al., 2012). Building on this line of work, FOOPS! extends pitfall detection with an explicit focus on FAIR principles, linking ontology quality assessment to broader concerns of findability, accessibility, interoperability, and reuse (Garijo et al., 2021). Together, these tools exemplify a shift from ad hoc ontology design toward continuous, tool-supported governance of schema quality. For evidence infrastructures, this governance dimension matters because schema defects propagate directly into user-facing breakdowns: ambiguous property names, missing labels, and inconsistent value types appear as confusing facets, brittle filters, and error-prone submission workflows.

From the perspective of evidence-synthesis infrastructures, these developments have two direct consequences. First, they underline that schemas and ontologies must be treated as living system components whose evolution, validation, and quality assurance are integral to platform operation rather than one-off design activities. Second, they suggest that user interfaces cannot be decoupled from schema logic: constraints and quality checks inevitably surface at the interaction layer, whether through form generation, validation

feedback, or submission rejection. In this sense, schema operationalisation is not merely a backend concern but a core part of the user experience. Maintaining predictable interaction therefore requires not only a stable conceptual model, but also interface-facing schema work: human-readable labels and definitions, stable facet/property mappings, and controlled term presentation that reduces vocabulary mismatch during filtering and entry.

For InterDev, which relies on ontology-aligned representations of evaluation evidence, this literature motivates a design in which structural constraints, vocabulary governance, and user-facing workflows are tightly integrated. Rather than assuming that a schema language alone guarantees quality, InterDev operationalises constraints through ontology-aligned interaction primitives (faceted filters and structured submission fields) combined with a consistent validation step prior to ingestion. This design treats the ontology as part of the interaction contract: it must support not only downstream querying and interoperability, but also usable, predictable filtering and contribution by non-expert users. Operationalisation therefore includes vocabulary work needed to make schema elements usable at the interface layer, such as stable labels, synonyms, and controlled term presentation for autocompletion and form generation. Without this interface-facing layer, constraint mechanisms can enforce validity while still leaving users unable to select or interpret the intended concepts, increasing both interaction cost and the risk of silent data quality failures.

3.1.4 Pillar 4: Knowledge graph interaction patterns for non-experts

Knowledge graphs provide structured representations of entities and relationships that support applications ranging from semantic search to scientific knowledge management (Heath & Bizer, 2011). However, the expressive power of RDF- and ontology-based models comes at the cost of increased abstraction, which poses substantial challenges for end-user interaction, particularly for users unfamiliar with formal query languages or graph-based data models. As a result, a large body of work has focused on interface patterns that make knowledge graphs accessible to non-expert users while preserving the benefits of structured representations.

Empirical studies underline the importance of user-centred design in this space. Obrezkov et al. (2023), in an evaluation of the Open Research Knowledge Graph, show that users struggle with navigation, the interpretation of ontological terms, and the cognitive burden of understanding graph structures. Earlier work by Dadzie and Rowe (2011) similarly

identifies the lack of appropriate visual metaphors and interaction techniques as a barrier to wider adoption of linked data applications. A related design response is to avoid “graph-first” interaction entirely. ORKG’s comparison-oriented views illustrate how KG structure can be surfaced through tabular and side-by-side representations that support sense-making and reduce the need for users to reason directly about triples or URIs (Obrezkov et al., 2023). This “graph-backed, not graph-first” stance is particularly relevant for evidence platforms, where users primarily need to locate, compare, and curate bounded sets of records rather than to inspect arbitrary graph topology.

Visual exploration remains a common strategy for mitigating graph complexity. Survey work by Katifori et al. (2007), Kriglstein (2009), and Dudáš et al. (2018) documents a wide range of techniques, with node-link diagrams and force-directed layouts among the most common. These representations provide intuitive navigational affordances and have been widely adopted to support exploratory browsing (Dadzie et al., 2011; Dudáš et al., 2018; Kriglstein, 2009). At the same time, empirical and survey evidence shows that node-link views scale poorly, becoming cluttered as graph size and density increase (Katifori et al., 2007). Controlled comparisons further indicate that alternative encodings can be superior for specific tasks. Ghoniem et al. (2004) show that matrix-based representations can outperform node-link diagrams for tasks involving dense connectivity, even if they are less intuitive for open-ended exploration. The design implication is that visual encodings should be selected and evaluated against the specific tasks an evidence platform must support, not treated as a universal default (Ghoniem et al., 2004; Katifori et al., 2007). In evidence-synthesis settings, this motivates prioritising interaction designs that reduce cognitive overhead for filtering, inspection, and comparison over designs that require users to interpret graph structure directly.

To address scalability and cognitive load, many systems combine lightweight exploration with filtering, faceting, and constrained query construction. Faceted search over RDF knowledge graphs allows users to iteratively narrow result sets using semantically meaningful dimensions, reducing both visual complexity and navigation overhead (Arenas et al., 2016). Query builders such as Sparklis and visual exploration tools such as Sparnatural pursue a similar goal by guiding users through controlled query construction using domain vocabularies and interaction constraints rather than requiring direct SPARQL

authoring (Ferré, 2016; Francart, 2023). Across these systems, progressive disclosure is operationalised through concrete micro-interactions such as semantic autocompletion, incremental expansion of the information shown, and interaction cues that help users anticipate the effect of a refinement step before committing to it (Arenas et al., 2016; Ferré, 2016; Francart, 2023; Katifori et al., 2007). For InterDev, these patterns motivate a faceted, task-oriented exploration model in which users iteratively constrain the evidence space and then curate selected impact evaluations into collections that act as bounded working sets for downstream comparison and export.

Natural language interfaces represent another strategy for lowering entry barriers. Liang et al. (2021) show how natural language querying can make knowledge graphs accessible to users who are unfamiliar with formal query languages. However, benchmark-based evaluations indicate that performance degrades as query complexity increases, particularly for compositional or multi-hop questions (Dubey et al., 2019). This limitation motivates hybrid interaction designs in which natural language is paired with mechanisms that constrain interpretation, guide users toward valid intents, or ground responses in explicit graph contexts. Work that combines language models with graph reasoning components can be read in this light, because it indicates that higher-level natural language access often requires structured reasoning support rather than purely surface-form parsing (Yasunaga et al., 2021). In evidence-centric systems, this suggests that natural language support is most defensible when it is bounded to an explicit evidence context (e.g., a user-curated subset) and when its outputs remain inspectable in terms of which records are being referenced.

Finally, interface patterns cannot be separated from backend access and performance constraints. Responsive progressive disclosure, interactive faceting, and bounded subgraph exploration depend on query and retrieval strategies that return partial results quickly and support incremental refinement. Lightweight access methods such as Triple Pattern Fragments were proposed in part to enable more responsive, web-friendly exploration by changing how query workload is distributed between servers and clients, with corresponding trade-offs in completeness, latency, and interaction smoothness (Verborgh et al., 2016). For evidence-centric systems, this coupling matters because interaction designs that encourage iterative refinement and verification are only viable if the system

can maintain predictable response times as the graph grows and if the semantics of filtering remain stable across interface iterations.

Taken together, prior work establishes a design space characterised by trade-offs between expressivity, scalability, and usability. Visualisation techniques, faceted exploration, query builders, and natural language interfaces each address different parts of this space, but none eliminate the tension between graph complexity and user comprehensibility. For InterDev, the key implication is that non-expert interaction should be organised around constrained, task-oriented exploration of bounded evidence subsets, supported by progressive disclosure and by backend strategies that keep these interactions responsive and inspectable (Arenas et al., 2012; Ferré, 2016; Francart, 2023; Verborgh et al., 2016). This motivates an interface approach in which filtering and inspection are exposed through user-facing facets and detailed views, while the knowledge graph remains the underlying representation that ensures interoperability, structured retrieval, and traceable curation.

3.1.5 Pillar 5: LLM and knowledge graph integration for extraction and retrieval

LLMs have shifted user expectations from form-based query interfaces toward conversational and generative interaction, but their adoption in evidence-centric systems raises specific technical and epistemic challenges. While generative fluency enables flexible access to complex information spaces, empirical evaluations show that LLMs remain prone to hallucinations, inconsistent reasoning, and sensitivity to prompt formulation, particularly in domain-specific or multi-step tasks (Bang et al., 2023). For applications that require traceability, auditability, and reproducibility, such as evidence synthesis and curation, these limitations make it insufficient to treat LLMs as standalone answer generators.

One response to these concerns has been the development of hybrid interface and system architectures that combine LLMs with external knowledge sources. Rather than relying solely on parametric model knowledge, retrieval-augmented generation approaches Lewis et al (2020) condition model outputs on dynamically retrieved documents or structured knowledge contexts, including knowledge graph fragments. This pattern seeks to improve factual grounding while retaining the flexibility of natural language interaction. In the

biomedical domain, KRAGEN exemplifies this approach by integrating knowledge graph retrieval with prompt-based generation, yielding improvements in factual accuracy and answer reliability relative to purely generative baselines (Matsumoto et al., 2024). More broadly, such architectures decouple knowledge maintenance from model training, allowing underlying knowledge bases to be updated without retraining the language model.

Beyond single-step retrieval, several systems explore tighter forms of coupling between LLMs and structured knowledge. QA-GNN integrates graph neural reasoning with language model representations to support multi-hop question answering over knowledge graphs, demonstrating that structured relational information can guide and constrain model inference (Yasunaga et al., 2021). UNIQORN extends this idea by dynamically constructing context graphs from both RDF knowledge bases and unstructured text, and using these graphs to steer retrieval and reasoning during question answering (Pramanik et al., 2024). These approaches illustrate a shift from treating retrieval as a simple pre-processing step toward viewing it as an iterative, model-guided process that incrementally shapes the evidence context.

Iterative reasoning architectures push this integration further by explicitly interleaving graph exploration and language model inference. The Tree-of-Traversals approach constructs progressively expanding subgraphs through a sequence of model-guided queries, enabling complex multi-hop reasoning over large knowledge graphs (Markowitz et al., 2024). While such methods show promise in handling compositional queries, they also introduce new challenges related to interaction latency, error propagation across reasoning steps, and the opacity of intermediate decision processes. From a system design perspective, these trade-offs highlight the tension between expressive reasoning capabilities and the need for predictable, inspectable behaviour in user-facing applications.

At the interface level, these architectural developments have been accompanied by a move away from minimal chatbot designs toward more structured interaction approaches. Studies of LLM interfaces show that purely conversational layouts are often insufficient for high-stakes or high-cognitive-load tasks, motivating the inclusion of retrieval controls, output comparisons, and guided prompting mechanisms (Bang et al., 2023; Reynolds & McDonell, 2021) (Bang et al., 2023; Reynolds & McDonell, 2021). Prompt scaffolding in particular has been shown to improve output consistency and reduce user frustration by constraining the space of possible interactions and making model behaviour more

predictable (Reynolds & McDonell, 2021). These findings align with the broader pattern observed in knowledge graph interfaces, where guided interaction and progressive constraint tend to outperform unrestricted exploration for complex tasks.

For evidence-centric platforms, the central implication of this literature is that LLMs should be treated as components within controlled evidence pipelines, rather than as autonomous sources of answers. Retrieval-augmented and graph-guided architectures provide mechanisms for grounding model outputs in structured evidence, but they do not by themselves guarantee transparency or correctness. Instead, they shift the design problem toward (i) how evidence context is selected and bounded, (ii) how provenance and intermediate steps are surfaced in ways that remain usable to non-experts, and (iii) how interfaces support verification and correction rather than passive acceptance. This framing maps directly onto InterDev’s design, which evaluates LLMs in two complementary roles: as an ingestion-time component that assists ontology-aligned extraction into the knowledge graph under validation and human review, and as a use-time interaction layer that mediates exploration and synthesis over user-curated, bounded knowledge graph subsets. In both roles, the emphasis is on constraining model behaviour through schema-aligned representations and user-facing controls so that the benefits of natural language interaction are obtained without sacrificing auditability and reproducibility.

3.1.6 Pillar 6: Human–AI interaction for appropriate reliance

As LLM-based systems become embedded in professional and research workflows, questions of usability are increasingly intertwined with questions of trust, confidence, and appropriate reliance. Unlike traditional decision-support tools, LLMs produce fluent outputs that can create an impression of competence and authority even when underlying content is uncertain or incorrect. This creates an interaction challenge that is distinct from “acceptance”: not simply whether users trust the system, but whether their reliance is calibrated to the system’s actual reliability in the task at hand.

Empirical studies highlight the fragility of this calibration. Nourani et al. (2019) show that the presence and form of explanations can influence user trust without necessarily improving decision quality. Explanations that appear meaningful but lack diagnostic value may increase confidence without increasing accuracy, and can even mislead users. This

complicates the common assumption that adding transparency cues automatically improves human–AI collaboration. Instead, explanation interfaces should be evaluated for their behavioural effects on judgement quality, error detection, and verification practices, not only for perceived usefulness.

Recent work on LLM behaviour reinforces these concerns. Simhi et al. (2025) demonstrate that language models can produce highly confident yet incorrect outputs, and that users may judge such responses as more helpful or credible than more cautious but accurate ones. In evidence-centric tasks, the resulting risk is overreliance: users may accept outputs that are wrong but presented with high apparent certainty. The core failure mode is therefore not merely model error, but interaction designs that do not sufficiently prompt users to question, verify, or contextualise model outputs.

These findings shift the design objective from maximising trust to supporting appropriate reliance. The goal is not to make systems appear more convincing, but to help users form accurate mental models of system competence and limitations, and to scaffold verification and critical engagement where warranted. For evidence-centric platforms, this reframing has direct implications for interface features such as citations, provenance indicators, confidence cues, and alternative views of answers. Such features can function either as epistemic supports or as persuasive signals depending on how they are presented. Their effects are therefore not guaranteed: provenance cues may reassure users, but they may also increase scepticism or redirect effort toward checking, and the net impact on confidence and calibration must be treated as an empirical question.

For InterDev, which integrates LLM-mediated interaction into workflows for evidence exploration, curation, and submission, these considerations are particularly salient. InterDev is not designed to replace expert judgement but to support it by reducing interactional burden while preserving user control over evidence selection and contribution. Concretely, this motivates interaction designs that keep LLM assistance bounded to explicit evidence contexts (e.g., curated collections) and that surface traceable links between system outputs and the underlying evaluations when users are expected to rely on them. It also motivates treating “trust” as insufficient as a sole outcome measure. In InterDev’s evaluation programme, the question is not only whether LLM assistance improves perceived usability, but whether it changes workload, confidence, and verification behaviour in ways that preserve or improve judgement quality. Accordingly, evaluation

must extend beyond conventional usability instruments to include workload measures, task-level confidence elicitation, and calibration analyses that compare subjective confidence with objective task performance. In this way, the human–AI interaction literature, insofar as it concerns generative AI, provides both a cautionary backdrop and a methodological foundation for assessing whether InterDev supports not just efficient interaction, but responsible and well-calibrated use of automated assistance.

3.1.7 Summary: Design requirements and interaction principles

3.1.1–3.1.6 motivate InterDev’s core design stance: evidence platforms require interoperable representations, operationalised structure, and interaction patterns that enable non-experts to explore, curate, and reuse bounded evidence sets. In response, InterDev represents impact evaluation evidence as ERCT-aligned RDF in a knowledge-graph backend, and centralises access through backend-mediated query generation to preserve stable retrieval semantics without requiring users to author SPARQL queries.

To support reuse and inspection of ongoing work, InterDev treats user-defined collections as bounded working sets derived from the global evidence graph, forming the primary unit for comparison, export, and later synthesis-oriented interaction. Rather than assuming that schema formalisms alone guarantee quality, InterDev operationalises schema commitments at the interaction layer: interfaces expose ontology-aligned facets and structured submission fields, while the backend serialises and validates outputs prior to ingestion to protect the integrity of the shared evidence base. In the current implementation, this enforcement is realised through controlled field structures, vocabulary alignment for facets, and pre-ingestion validation gates (e.g., RDF parseability), with ontology governance changes (labels, facet properties, provenance descriptors) introduced iteratively in response to observed breakdowns.

Finally, LLM support is treated as mixed-initiative assistance rather than autonomous answer generation, and is evaluated in two complementary roles: ingestion-time extraction that pre-populates submission fields under human review (LLM-assisted submission), and use-time mediation that supports exploration and synthesis over user-curated collections (LLM-mediated exploration). Accordingly, InterDev’s evaluation programme treats usability improvements as necessary but insufficient, extending measurement to workload,

confidence, verification behaviour, and confidence–correctness calibration. This reflects the central lesson from the literature and the empirical iterations: grounding and provenance cues enable audit and checking, but their impact on reliance cannot be assumed and must be tested empirically.

Across the six pillars, a common set of design tensions emerges. Evidence platforms must balance expressive, interoperable representations against usability for non-expert users; strong schema control against flexible contribution and evolving evidence types; conversational fluency and LLM mediation against inspectability and predictable behaviour; and lower interactional effort against the risk of overreliance or weakened verification. InterDev is motivated by these tensions rather than by any assumption that they can be eliminated. Its design therefore combines bounded evidence contexts, ontology-aligned validation, provenance-aware interaction, and human-in-the-loop mediation as a way of negotiating these trade-offs in a form that can be implemented and evaluated empirically.

The six pillars define a design space: they identify what evidence platforms could do (representational choices, interaction patterns, and LLM integration architectures) and what risks those choices introduce, particularly around auditability and calibrated reliance. However, selecting among these options requires grounding in the concrete task demands of the target domain. InterDev is positioned as an evidence platform for international development synthesis workflows, where filtering, contextualisation, and cross-study comparison have distinctive recurring patterns. The next sections therefore shift from general state-of-the-art review to domain grounding, using a review-of-reviews to derive initial user requirements from documented systematic review practice.

The six pillars also constrain specific design decisions. Pillars 1 and 2 justified treating extracted evidence and curated subsets as structured, provenance-linked units rather than as transient interface state. Pillar 3 required that ontology commitments be operationalised through controlled fields, validation, and stable facet semantics rather than left as purely conceptual models. Pillar 4 constrained the interaction design toward faceted, task-oriented exploration and collection-based working sets instead of graph-first navigation or direct SPARQL authoring. Pillar 5 motivated the use of LLMs only within bounded, schema-aligned workflows for submission and exploration, rather than as unconstrained answer generators. Pillar 6, finally, required that evaluation extend beyond usability and efficiency

to include workload, confidence, verification behaviour, and calibration. Taken together, these constraints help define the architecture and interaction strategy realised in Chapter 4 and the evaluation logic developed in Chapter 5.

3.2 Domain grounding and requirements elicitation from systematic reviews

Figure 3 provides a focused view of Phases 1–2 of the six-phase User-Centred Design (UCD) framework introduced in Section 1.5. In the overall programme, these phases establish (i) an evidence-based understanding of user needs and contexts and (ii) an explicit set of initial domain-grounded user requirements that constrain subsequent iterations and evaluation.

In this thesis, Phases 1–2 are operationalised through a domain-grounding study based on a review-of-reviews in international development. Rather than relying on open-ended elicitation alone at the outset, the review-of-reviews (i.e., a review of systematic reviews) provides a structured account of how evidence synthesis is conducted in practice; how researchers filter, contextualise, extract, and synthesise evaluation evidence, and where breakdowns recur across sectors and review types. This approach supports early-stage requirements formation in two ways: it anchors requirements in documented research practice rather than assumed workflows, and it provides traceable justification for design choices that later iterations can test and refine.

The review-of-reviews is treated as an initial task and workflow analysis at the level of established evidence-synthesis practice. Findings from this analysis are reported in Section 3.2.4 as recurring themes based on user needs (Phase 1 outputs). These are then translated into initial system requirements in Section 3.2.5 (Phase 2 outputs), which form the baseline that later phases refine through prototyping and study-based evaluation (Section 1.5;

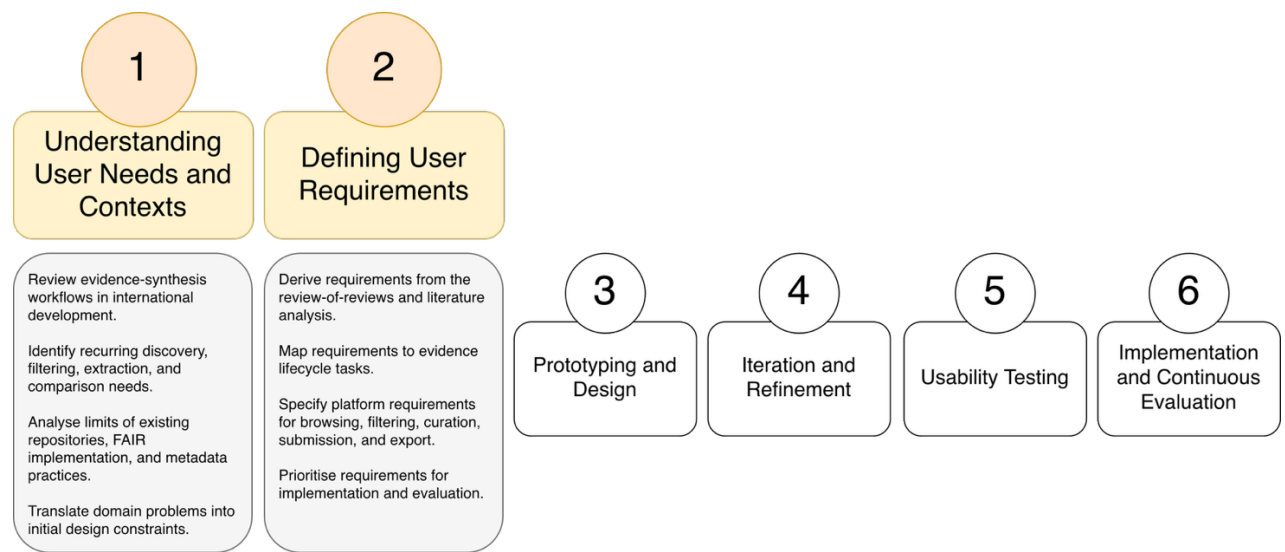


Figure 3: InterDev Methodological Process - Understanding User Needs and Contexts, Defining User Requirements

3.2.1 Understanding the Process: A Review of Systematic Reviews

This domain-grounding study treats systematic reviews and meta-analyses as empirical traces of synthesis practice in international development. Rather than eliciting requirements only through speculative workflow descriptions, it analyses how published reviews operationalise key steps, such as screening and filtering criteria, contextual coding, extraction of effect-level data, and synthesis choices, and where authors repeatedly report breakdowns (e.g., missing metadata, heterogeneous outcomes, limited access to underlying materials). The analysis spans major development sectors (education, health, social protection, and economic development) to ensure that requirements reflect cross-sector synthesis practice rather than a single domain’s conventions

These observed practices are then translated into initial InterDev requirements in Section 3.2.5, providing traceable justification for the design choices evaluated in Chapters 4–5.

3.2.2 Review Framework and Parameters

A key question for any systematic review is which review framework to implement in order to provide the review structure and replicability. For the purpose of this research, the SPIDER (Sample, Phenomenon of Interest, Design, Evaluation, Research Type) (Cooke et

al., 2012) framework was chosen for this review-of-reviews is based on its suitability for qualitative and mixed-methods research. Traditional systematic reviews in evidence-based fields have relied heavily on the PICO (Population, Intervention, Comparison, Outcome) framework, initially developed for clinical research (Richardson et al., 1995). While PICO is highly effective for structuring research questions in quantitative studies, it is less appropriate for qualitative research, which often does not involve a direct intervention or comparison group. As outlined by Cooke, Smith, and Booth (2012), the SPIDER framework provides a more nuanced approach for reviews synthesizing qualitative and mixed-method evidence. Unlike PICO, which focuses on interventional studies, SPIDER emphasizes methodological rigor, study design, and research context, making it particularly relevant for the evaluation of systematic reviews in international development. Given that an important facet of this study is examining how evaluation data is filtered, contextualized, and synthesized in development research, SPIDER offers a structured yet flexible framework that aligns with the complexity of policy-oriented systematic reviews. The following sections outline the research parameters using this approach.

Methley et al. (2014) further compared PICO, PICOS, and SPIDER to assess their effectiveness in searching for qualitative research. Their findings highlight that while PICO yields a higher number of search results, it often lacks specificity in qualitative evidence synthesis. SPIDER, on the other hand, demonstrates higher specificity, making it more appropriate when searching for and synthesizing qualitative and mixed-method studies. Given that the study for this thesis research examines how evaluation data is filtered, contextualized, and synthesized in development research, the SPIDER framework offers a structured yet flexible approach that aligns with the complexity of policy-oriented systematic reviews. The following sections outline the research parameters using the SPIDER framework for the study.

3.2.2.1 Sample

This review-of-reviews of systematic reviews focuses on studies that assess evaluation methodologies within the field of international development. The participants in this research are systematic reviews that examine evaluation processes relevant to researchers and practitioners working in development contexts. These systematic reviews explore intervention effectiveness in key sectors, including but not limited to education, social

protection, public health, and economic development. Education-focused reviews may assess interventions aimed at improving learning outcomes and school attendance, while social protection reviews often examine conditional and unconditional cash transfer programs. Reviews in public health may evaluate community health initiatives and maternal health interventions, whereas those in economic development may study the impact of microfinance, employment programs, and infrastructure projects. The included systematic reviews must focus on how evaluation data is filtered, contextualized, and synthesized to enhance its relevance for policy and research.

3.2.2.2 Phenomenon of Interest

The central focus of this study is to examine how systematic reviews within international development research manage key aspects of evaluation methodologies. A particular emphasis is placed on the filtering mechanisms used to select studies, with priority often given to randomized controlled trials, quasi-experimental designs, or mixed-method approaches. Contextualization processes are also a key area of investigation, as systematic reviews frequently account for socio-economic and geographical factors, such as regional classifications, to ensure policy relevance. The study further examines the synthesis approaches employed by systematic reviews, including meta-analysis, narrative synthesis, and mixed-methods integration. Additionally, it explores the challenges involved in cross-study comparisons, particularly regarding the harmonization of heterogeneous evaluation data. Finally, the research assesses the impact of systematic reviews on policymaking, determining the extent to which they contribute to shaping development policies and informing decision-making processes.

3.2.2.3 Design

The study undertaken for this thesis exclusively includes systematic reviews and meta-analyses that follow a structured and transparent methodology. To be eligible for inclusion, systematic reviews must explicitly state their search strategies and inclusion/exclusion criteria. They must also employ explicit synthesis methods, whether quantitative, qualitative, or mixed-methods, and apply critical appraisal tools to assess the quality of included studies. Reviews that rely solely on narrative discussion without systematic methodology are excluded from this study.

3.2.2.4 Evaluation

The evaluation component of this study focuses on analysing key methodological elements in systematic reviews. Particular attention is given to the filtering mechanisms used to determine study inclusion based on methodology, region, or policy relevance. Additionally, contextualization strategies are assessed to understand how systematic reviews align their findings with real-world development scenarios, such as through the socioeconomic stratification of results. The study also evaluates the synthesis approaches utilized in systematic reviews, including meta-analysis, qualitative thematic analysis, and Bayesian synthesis. Further, it identifies the challenges encountered in the review process, including variations in reporting standards and data heterogeneity, which can hinder cross-study comparability. Lastly, the study examines the utility of systematic reviews in policymaking by assessing the extent to which they provide actionable insights for development practitioners.

3.2.2.5 Research Type

The evidence base targeted by this review is methodologically heterogeneous, reflecting the mixed nature of international development evaluation practice. Included reviews span quantitative meta-analyses that pool standardised effects or compute comparable impact estimates, qualitative evidence syntheses that derive themes about mechanisms and implementation, and mixed-method reviews that combine statistical aggregation with narrative explanation of context and pathways. This choice is deliberate: InterDev is intended to support research synthesis across designs rather than to optimise for a single study type. Accordingly, the review-of-reviews treats “research type” as a first-class dimension of eligibility and coding, ensuring that requirements derived from the corpus reflect the needs of users who must traverse between experimental evidence, quasi-experimental studies, and qualitative or theory-driven accounts when constructing policy-relevant conclusions.

3.2.3 Search Protocol and Strategy

Consistent with the SPIDER framework, the search protocol for the discovery and inclusion of papers in this review is outlined here. For this review, keywords and Boolean operators were identified based on scoping searches and previous research (Murtagh-White, 2021)

and established databases for research and review documents, both academic and institutional, were included.

3.2.3.1 Databases and Repositories Searched

Below is a comprehensive list of academic and institutional databases used in this review. Databases repositories that offer systematic reviews on any theme in international development were included that offer some form of search functionality.

Database/Repository	Reason for Inclusion	Type
Scopus	Comprehensive academic database covering multidisciplinary research, including development studies and policy evaluations.	Academic Database
Web of Science (WoS)	High-impact, peer-reviewed literature, ensuring quality systematic reviews in international development.	Academic Database
PubMed	Covers health-related systematic reviews, particularly relevant for public health interventions in LMICs.	Academic Database
EconLit	Specialized database for economic development research, including impact evaluations and systematic reviews.	Academic Database
ERIC	Focused on education research, including systematic reviews on learning interventions and school policies.	Academic Database
JSTOR	Contains historical and policy-related systematic reviews in social sciences and development studies.	Academic Database
Google Scholar	Captures grey literature and recent systematic reviews not yet indexed in traditional databases. Used as a supplementary source.	Academic Database
3ie (International Initiative for Impact Evaluation) Systematic Review Database	Repository of systematic reviews and meta-analyses focused on international development, impact evaluations, and policy interventions.	Institutional Repository
Campbell Collaboration Library of Systematic Reviews	Contains systematic reviews on social and economic interventions, emphasizing rigor in development research.	Institutional Repository

Database/Repository	Reason for Inclusion	Type
Cochrane Library	Source of high-quality systematic reviews, particularly useful for public health interventions in development settings.	Institutional Repository
World Bank Open Knowledge Repository	Contains impact evaluations and systematic reviews conducted by the World Bank on development policies and programs.	Institutional Repository
UNICEF Evidence Synthesis Repository	Houses systematic reviews related to child-focused development programs, education, and health interventions.	Institutional Repository
OECD Development Library	Contains policy-focused systematic reviews on governance, economic development, and social protection in LMICs.	Institutional Repository
USAID Development Experience Clearinghouse (DEC)	Now defunct ³ repository of systematic reviews and evaluations conducted for USAID-funded international development programs.	Institutional Repository

Table 3: Repositories and Databases used for Review of Systematic Reviews

3.2.3.2 Keyword Strategy

Where possible, a Boolean search strategy was used across databases to ensure comprehensive retrieval of systematic reviews. The following search strings were employed:

```
("systematic review" OR "meta-analysis" OR "evidence
synthesis" OR "realist review" OR "scoping review")
AND
("international development" OR "development
evaluation" OR "impact evaluation" OR "policy
evaluation")
AND
("education" OR "health" OR "social protection" OR
"economic development" OR "poverty reduction" OR
```

³ <https://usaid.onl/>

```
"livelihoods" OR "cash transfer" OR "agriculture" OR
"public health" OR "governance" OR "gender
empowerment")
AND
("LMIC" OR "low-income countries" OR "middle-income
countries" OR "developing countries")
AND
("effectiveness" OR "impact" OR "outcomes" OR "cost-
effectiveness" OR "intervention")
```

While this search string was followed as closely as possible for all repositories and databases, a certain amount of customisation was necessary in order to satisfy database specific requirements. For example, Scopus and Web of Science followed the format:

```
TITLE-ABS-KEY ("systematic review" OR "meta-analysis")
AND ("international development" OR "policy
evaluation")
```

Whereas Pubmed followed the format:

```
("systematic review"[Publication Type] OR "meta-
analysis"[Publication Type]) AND ("global health" OR "low-
income countries")
```

In addition to relevant works discovered through these database searches, backward and forward citation searching, whereby references of included systematic reviews and newer reviews that cited included systematic reviews were included (Booth et al., 2021; Greenhalgh & Peacock, 2005).

3.2.3.3 Selection Criteria

The following table outlines the selection criteria used to determine the eligibility of systematic reviews included in this study. These criteria ensure that only high-quality, methodologically rigorous reviews relevant to international development evaluations in Low- and Middle-Income Countries (LMICs) are considered.

Criterion	Description
Study Type	Must be a systematic review, meta-analysis, evidence synthesis, realist review, or scoping review. Narrative literature reviews or single empirical studies without a systematic review methodology are excluded.
Timeframe	Studies published between 2005 and 2024 to ensure the review captures recent and relevant evidence from the past 20 years.
Geographic Scope	Studies must focus on Low- and Middle-Income Countries (LMICs) as classified by the World Bank. Reviews that include both high-income and LMIC contexts are considered if findings are disaggregated.
Methodological Rigor	Systematic reviews must explicitly report search strategies, inclusion/exclusion criteria, and synthesis methods (e.g., meta-analysis, narrative synthesis, realist synthesis).
Publication Type	Must be peer-reviewed journal articles or systematic reviews published by recognized development institutions (e.g., World Bank, 3ie, DFID, Cochrane, Campbell Collaboration).
Language	Only studies published in English were included due to feasibility constraints.

Table 4: Selection Criteria for Review of Systematic Reviews

The result of this procedure is a corpus of reviews that explicitly report inclusion criteria, contextual coding decisions, and synthesis methods. The analysis in Section 3.2.4 codes these reviews for recurring task demands and breakdowns that an evidence platform must support. To keep the reasoning traceable, each identified theme is linked to the review practices that instantiate it, and each resulting requirement is mapped back to those themes.

3.2.4 Findings from Review-of-Reviews: Understanding User Needs and Contexts

This section derives InterDev’s initial needs and task demands from a review-of-reviews in international development, treating systematic reviews and meta-analyses as empirical traces of how evidence is actually filtered, coded, synthesised, and reported in practice. This analysis investigates recurring workflow decisions reported across influential reviews

in education, health, social protection, and economic development. Four themes were surfaced consistently in this corpus, reflecting recurrent workflow demands: (i) filtering and contextualisation of evidence, (ii) data extraction and synthesis practices, (iii) analytic focus on intervention impacts and mechanisms, and (iv) structural and contextual challenges to cross-study comparison. From these themes, three requirements were identified: the ability to filter evidence by methodological rigor and policy-relevant context (R1) (Baird et al., 2014; Lagarde et al., 2007; Pega et al., 2022; Bastagli et al., 2016; Evans & Popova, 2016), the ability to move between granular effect-level data and higher-level narrative summaries depending on user role and task (R2) (Fryer, 2017; McGuire et al., 2022; Millán et al., 2019; Snilstveit et al., 2015), and mechanisms to harmonise heterogeneous outcomes and study designs to support cross-study comparison and mixed-method synthesis (R3) (Van Rooyen et al., 2012; Waddington et al., 2012; Masino & Niño-Zarazúa, 2016; Akyeampong et al., 2023). The remainder of the section unpacks these themes and shows how they translate into requirements. Section 3.2.5 then reformulates these themes into explicit system requirements that constrain the design space identified in Section 3.1.

3.2.4.1 Filtering and Contextualisation of Data

A key theme across the systematic reviews is the need for rigorous filtering mechanisms to ensure the relevance and credibility of included studies. The reviewed works frequently prioritized methodological rigor, often impact evaluations and quasi-experimental designs as the gold standard for causal inference (Baird et al., 2013; Lagarde et al., 2007; Pega et al., 2022). However, as multiple reviews noted, qualitative and mixed-method studies are also essential for understanding the mechanisms and contextual factors driving development outcomes (Van Rooyen et al., 2012; Waddington et al., 2012).

Systematic reviews also commonly apply contextual filters, such as geographic, economic, and policy-relevant classifications. For example, Fryer (2017) used the Human Development Index (HDI) to classify studies by income level, ensuring that findings were relevant to specific socioeconomic contexts. Similarly, Bastagli et al. (2016) categorised evaluations of cash transfers by their policy impact, focusing on domains such as education, financial savings, and economic well-being. Other reviews emphasized the importance of sector-specific filtering. For instance, education reviews grouped interventions by age

group (early childhood, school-based, adult learning) and pedagogical approach (Akyeampong et al., 2023; Conn, 2017; Evans & Popova, 2016; Ganimian & Murnane, 2016; Glewwe et al., 2011; Kremer et al., 2013; Krishnaratne & White, 2013; McEwan, 2015; Rodriguez-Segura, 2022).

These findings represent a core user need: tools that allow researchers to filter evaluation data by methodology, regional relevance, and intervention type. InterDev must enable users to select studies not only based on the rigor of their evaluation design but also by contextual parameters, such as country income classification, policy focus, and intervention typology.

3.2.4.2 Data Collection Practices

The systematic reviews analysed reveal a wide diversity of data collection and synthesis approaches, reflecting the varying goals of researchers in international development. Reviews such as Bastagli et al. (2016) emphasize the importance of qualitative synthesis, particularly for capturing policy insights and implementation challenges. These approaches cater to policy-focused users who prioritize actionable, narrative-driven insights over technical statistical outputs.

Conversely, reviews like Fryer (2017) and Baird et al. (2014) rely on quantitative meta-analysis, extracting treatment effects from results tables to estimate pooled impact sizes. Meta-analyses are necessary for evidence aggregation and cross-study comparisons, allowing researchers to quantify intervention effectiveness across multiple settings (McGuire et al., 2022). However, systematic reviews frequently note the challenge of heterogeneous study designs. For example, education interventions often lack standardized learning outcome measures, complicating synthesis efforts (Snilstveit et al., 2015).

For a platform like InterDev, this diversity in synthesis approaches shows the need for structured metadata schemas that accommodate both granular data extraction (for meta-analysis) and high-level thematic synthesis (for qualitative insights). The ability to integrate quantitative treatment effects, policy lessons, and contextual factors within a single platform would significantly enhance its usability for diverse researcher needs.

3.2.4.3 Focus on Intervention Impact

A consistent finding across systematic reviews is that understanding intervention impact requires both outcome-oriented and mechanism-oriented approaches. Traditional reviews often focus on whether an intervention “works”. For example, systematic meta-analyses confirm that cash transfers reliably increase school enrolment, improve food security, and enhance household economic stability (Baird et al., 2014; Pega et al., 2022). However, more recent reviews incorporate mechanistic insights, exploring why certain interventions succeed or fail across contexts (Vaessen et al., 2020).

For instance, Fryer (2017) categorizes educational interventions into early childhood, school-based, and home-based models, allowing for comparative assessment of which approaches yield the strongest learning gains. Similarly, Bastagli et al. (2016) examine cross-sectoral effects of cash transfers, revealing that while these programs improve financial well-being and service utilization, their effects on long-term economic mobility and cognitive development are weaker (Millán et al., 2019).

Furthermore, newer reviews emphasize context-dependent outcomes. For example, the effectiveness of vocational training programs varies significantly based on labour market conditions and economic opportunities (Stöterau et al., 2022), a factor that was also highlighted in some older reviews (Tripney & Hombrados, 2013). This need for mechanism-sensitive interpretation and comparison strengthens R2 and R3 by requiring both effect-level access and structured context for cross-study interpretation.

These findings demonstrate the need for evaluation platforms to move beyond binary “effectiveness” conclusions and instead provide users with nuanced, contextualized intervention impact data. InterDev must incorporate interactive tools for comparing intervention effects across policy domains, socioeconomic settings, and target populations.

3.2.4.4 Contextual Challenges

Despite advances in systematic review methodologies, multiple reviews highlight persistent challenges in evaluation harmonization. A major obstacle is heterogeneity in methodologies, indicators, and reported outcomes. Many reviews note that cross-study comparability is limited by inconsistent measurement standards, particularly in education,

economic development, and multi-sector interventions (Masino & Niño-Zarazúa, 2016; McEwan, 2015).

Additionally, systematic reviews face gaps in long-term follow-up data, making it difficult to assess whether observed impacts are sustained over time. For example, while cash transfers demonstrably increase school attendance, long-term effects on cognitive skills and labour market success remain uncertain due to limited longitudinal studies (Millán et al., 2019). These gaps create challenges for researchers trying to derive generalizable policy lessons.

Moreover, reviews highlight resource constraints and accessibility barriers in evaluation research. Limited access to comprehensive datasets, especially in low-income and fragile state contexts, restricts the scope of systematic syntheses (Stewart et al., 2012). Many impact evaluations are produced as grey literature (e.g., NGO and government reports), making them difficult to systematically integrate into academic reviews. These constraints directly motivate R3, and they also inform R2 by clarifying why summary outputs must preserve traceable links to heterogeneous underlying evidence.

Addressing these challenges requires platforms that not only standardize and integrate data but also facilitate access to non-traditional evaluation sources.

3.2.5 Identification of Initial User Requirements

Building on the insights into user needs and contexts outlined in Section 3.2.4, the next step was to formalize a set of initial domain-grounded user requirements. These requirements served as the foundational guide for the design and development of the InterDev platform, ensuring that it aligns with the workflows, goals, and challenges of researchers in international development. Drawing from the systematic review findings and informed by principles of user-centred design (UCD), these requirements prioritize usability, data accessibility, and contextual relevance. Requirements were prioritised according to three criteria: recurrence across the review corpus, direct relevance to the core evidence-synthesis tasks identified in Section 3.2.4, and feasibility for implementation and evaluation within the initial InterDev design cycle. As will be seen later in the thesis, the requirements evolved iteratively in response to feedback generated during each evaluation phase of the framework.

3.2.5.1 User Requirement 1: Advanced Filtering and Contextualisation

A key requirement emerging from the systematic review is the need for robust filtering and contextualization capabilities to enhance the relevance and applicability of evaluation data. As systematic reviews in development research have demonstrated, filtering mechanisms allow researchers to identify relevant studies based on methodology, regional focus, and policy domain (Baird et al., 2014; Lagarde et al., 2007; Pega et al., 2022).

Filtering based on methodological rigor is important, as many reviews impact evaluations and quasi-experimental studies for causal inference (Lagarde et al., 2007; Waddington et al., 2012). However, systematic reviews also highlight the value of qualitative and mixed-methods studies in providing deeper contextual insights (Masino & Niño-Zarazúa, 2016; McGuire et al., 2022). Therefore, the InterDev platform must allow users to filter studies by evaluation design, ranging from experimental and quasi-experimental methods to qualitative and participatory research.

Filtering based on socioeconomic and geographic factors has also been highlighted as a need. Many systematic reviews classify evaluations using regional or income-level categories, such as the Human Development Index (HDI) or the World Bank's country income classifications (Fryer Jr, 2017; Snilstveit et al., 2015). Similarly, reviews examining cash transfer impacts across different policy domains (Bastagli et al., 2016) highlight the need to filter studies based on sectoral relevance (education, health, economic development, social protection, etc.).

Additionally, contextual filtering must address intervention typology and program design features. For instance, systematic reviews on education interventions classify studies based on pedagogical approach, age group, and learning outcome focus (Evans & Popova, 2016), while those on cash transfers differentiate between conditional and unconditional models (Baird et al., 2014). InterDev must enable users to apply these fine-grained filters dynamically.

Finally, usability considerations for this dimension are essential. Filtering tools must be intuitive and accessible to diverse users, including policymakers, practitioners, and researchers with varying levels of technical expertise. As systematic reviews emphasize the

challenges of accessing and synthesizing evaluation data (Mallett et al., 2012), InterDev must integrate pre-set filtering templates while also allowing users to create custom queries. By embedding such filtering mechanisms, InterDev can ensure that users can align evaluation data with their specific research and policy needs.

3.2.5.2 User Requirement 2: Access to Granular and Aggregated Data

Systematic reviews underscore a dual need for both granular and synthesised data access, as different user groups require different levels of analysis (Baird et al., 2014; Millán et al., 2019). Researchers often need granular datasets to perform meta-analysis, statistical modelling, and causal inference, while policymakers and development practitioners require high-level syntheses and policy insights.

Granular data access is essential for enabling quantitative meta-analysis and cross-study comparisons. Reviews such as Fryer (2017) and McGuire et al. (2022) extract treatment effects from experimental studies, pooling effect sizes to assess the overall impact of interventions. However, systematic reviews frequently report challenges due to heterogeneous outcome measures, inconsistent reporting standards, and missing data (Snilstveit et al., 2015). InterDev must support capture and access to effect-level data where reported, including effect sizes, uncertainty measures, and sample characteristics, to enable downstream meta-analysis.

Conversely, aggregated data access or summaries is important for users who require high-level summaries of findings across multiple studies. Reviews such as Bastagli et al. (2016) and Pega et al. (2022) focus on synthesizing policy implications rather than raw statistical outputs, catering to stakeholders who prioritize narrative-driven evidence. InterDev should support generation of aggregated summaries, with traceable links back to the underlying extracted evidence to support verification, using natural language processing (NLP) to extract key insights from systematic reviews where possible.

3.2.5.3 User Requirement 3: Tools for Cross-Study Comparisons and Harmonisation

A major challenge identified in the systematic review is the heterogeneity of evaluation data, which complicates cross-study comparisons and meta-analyses (Snilstveit et al., 2015; Van Rooyen et al., 2012). Researchers often struggle to synthesize findings due to variations in study designs, outcome measures, and intervention contexts (Millán et al.,

2019). To address this, InterDev must provide tools for harmonising and standardising evaluation data.

Systematic reviews emphasize the need for standardised metadata schemas to ensure consistent representation of interventions and outcomes. For instance, Waddington et al. (2012) advocate for using structured coding frameworks to classify studies, a practice increasingly adopted in systematic reviews (Akyeampong et al., 2023). The InterDev platform should integrate predefined ontologies, such as the ERCT ontology for impact evaluations, and must support controlled extension mechanisms to accommodate emerging evaluation methodologies without breaking interoperability.

Cross-study comparison tools must also facilitate multi-method synthesis, as development evaluations often include both quantitative and qualitative evidence (Masino & Niño-Zarazúa, 2016; McGuire et al., 2022). For example, systematic reviews on education interventions incorporate both meta-analytical effect sizes and qualitative insights on classroom implementation (Evans & Popova, 2016).

Additionally, usability considerations must be made throughout development. Many researchers lack expertise in semantic web technologies, making manual data standardization impractical (Hoefler, 2013; Kuric et al., 2019; Mallett et al., 2012). The InterDev platform will investigate abstracting complexity by offering automated tools for exploration and contributing data to the knowledge graph. LLMs have made breakthroughs in simplifying the summarisation and collation of data through RAG (P. Lewis et al., 2020); the potential of LLMs to assist users of non-technical backgrounds will be investigated.

Together, the user requirements R1–R3 (summarised in Table 5) specify the minimum interaction and representation capabilities that InterDev must provide: filtering and contextualisation over structured metadata, access pathways that support both effect-level inspection and narrative synthesis, and harmonisation support for cross-study comparison across heterogeneous designs. These requirements constrain the implementation choices described in Chapter 4 and motivate the evaluation focus in Chapter 5, including the emphasis on bounded evidence contexts, traceable provenance links, and reliance-sensitive measures (workload, confidence, verification behaviour, and calibration).

In implementation terms, the user requirements imply three concrete system commitments. First, R1 requires an exploration interface whose filtering operations are backed by stable,

ontology-aligned metadata fields and predictable query semantics. Second, R2 requires that individual records and curated subsets support both inspection at the level of evaluation metadata and export of analysis-ready representations suitable for downstream synthesis. Third, R3 requires that the platform make cross-study comparison feasible by enforcing a shared schema at ingestion time and by representing curated working sets as explicit, traceable units. Chapter 4 describes how these commitments are realised in InterDev’s architecture, and how iteration-specific extensions selectively manipulate submission and exploration modalities while holding the underlying evidence base and retrieval logic constant.

Requirement	Purpose	Key Features	Supporting Evidence
R1. Advanced Filtering and Contextualisation	Enable users to identify and access relevant studies based on methodological, geographic, socioeconomic, and policy criteria.	Dynamic filters based on: - Evaluation design (RCTs, quasi-experiments, qualitative) - Geographic and socioeconomic classifications (e.g., HDI, World Bank categories) - Policy domain and intervention typology - Usability: intuitive interfaces, pre-set templates, and custom queries	Baird et al. (2013, 2014); Lagarde et al. (2007); Pega et al. (2022); Bastagli et al. (2016); Evans & Popova (2016); Fryer (2017); Akyeampong et al. (2023); Conn (2017); Ganimian & Murnane (2016); Glewwe et al. (2011); Kremer et al. (2013); Krishnaratne & White (2013); McEwan (2015); Rodriguez-Segura (2022); Mallett et al. (2012)
R2. Access to Granular and Aggregated Data	Support different user groups (researchers, policymakers, practitioners) by enabling both in-depth analysis and high-level summaries.	- Access to granular data (e.g., effect sizes, confidence intervals, sample data) - Integration of structured metadata standards - NLP-based automated summary generation for narrative synthesis	Fryer (2017); McGuire et al. (2022); Millán et al. (2019); Snilstveit et al. (2015); Bastagli et al. (2016); Pega et al. (2022); Vaessen et al. (2020); Lewis et al. (2020); Stewart et al. (2012)
R3. Tools for Cross-Study Comparisons and Harmonisation	Address data heterogeneity and enable multi-method synthesis across diverse evaluation designs.	- Standardised metadata schemas and ontologies (e.g., ERCT) - Support for custom metadata extensions - Multi-method synthesis tools combining quantitative and qualitative data - Automated harmonisation via semantic tools and LLMs (e.g., RAG)	Snilstveit et al. (2015); Van Rooyen et al. (2012); Waddington et al. (2012); Masino & Niño-Zarazúa (2016); Millán et al. (2019); Akyeampong et al. (2023); Evans & Popova (2016); McGuire et al. (2022); Hoefler (2013); Kuric et al. (2019); Lewis et al. (2020); Stöterau et al. (2022); Tripney & Hombrados (2013)

Table 5: Initial User Requirements identified

Chapter 4

Implementation

This chapter describes how InterDev operationalises the requirements derived in Chapter 3 as a concrete linked-data platform. The emphasis is on explaining the implementation choices that are necessary to (i) support ontology-aligned filtering and contextualisation (R1), (ii) enable inspection, curation, and export at both record and collection level (R2), and (iii) make cross-study comparison and harmonisation feasible through schema-constrained ingestion and traceable working sets (R3).

InterDev is implemented as a client–server system backed by an ERCT-aligned knowledge graph. Across evaluation iterations, the evidence base and backend retrieval semantics are held stable, while interaction techniques and mediation layers (manual vs LLM-assisted submission; direct faceted exploration vs LLM-mediated exploration; provenance cue visibility) are varied. This separation is deliberate: it ensures that observed differences in usability, workload, confidence, and verification behaviour (Chapter 5) can be attributed primarily to interface and mediation design rather than to changes in the underlying corpus or query logic.

4.1 Design Approach and User-Centred Process

InterDev was developed using a six-phase User-Centred Design (UCD) process introduced in Section 1.5. This chapter focuses on the instantiation of Phase 3 (Prototyping and Design), where requirements are translated into concrete interaction artefacts and implementation choices. Phases 1–2, understanding user needs and defining requirements, provide the inputs to this phase, as outlined in section 3.2.

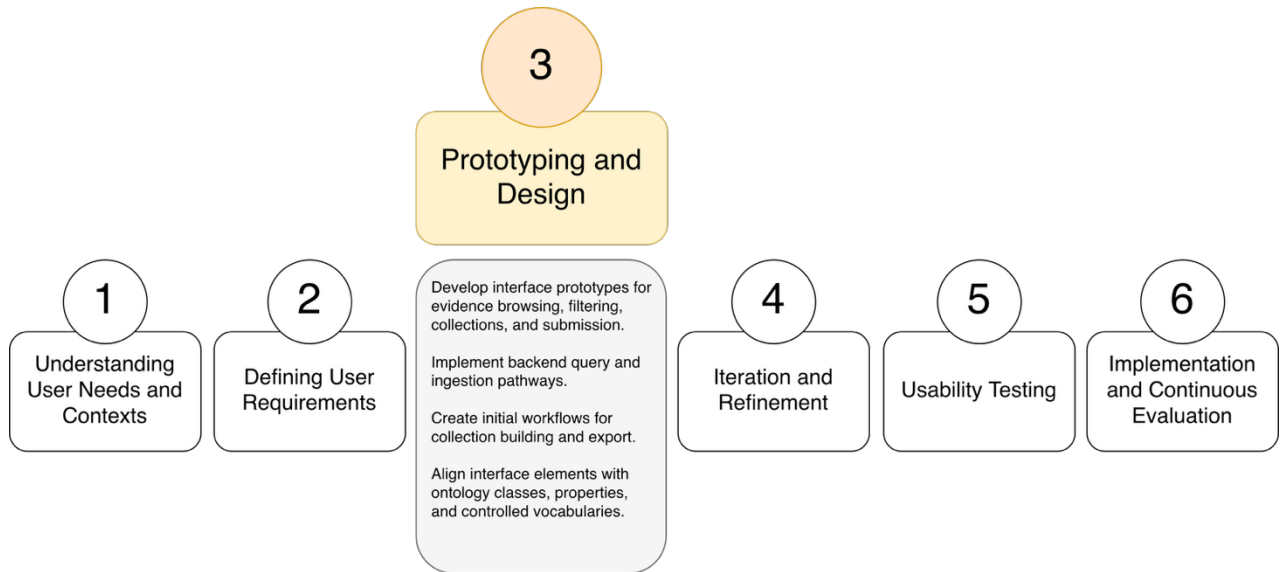


Figure 4: InterDev Methodological Process - Prototyping and Design

Figure 4 provides a focused view of Phase 3 within the broader UCD framework. In this thesis, “prototyping and design” refers to the progression from requirements to testable interface concepts (e.g., early mock-ups and functional prototypes) used to validate assumptions about evidence discovery, filtering, curation, export, and contribution before committing to full implementation. The canonical role of Phase 3 within the overall framework is described in Section 1.5; the remainder of this chapter documents the specific design artefacts produced, the system abstractions they motivated, and how these were implemented as the stable baseline for later evaluation iterations.

The remainder of the chapter therefore proceeds from infrastructure to interaction: it first describes the architecture and data layer that enforce stable retrieval and validation, then introduces the interaction abstractions through which users enact R1–R3, and finally summarises how iteration-specific extensions instrument those abstractions to test RO2–RO3.

For clarity, this chapter distinguishes between the core platform and the experimental layer. The core platform comprises the stable infrastructure held constant across evaluation iterations, including the ERCT-aligned knowledge graph, backend-mediated retrieval pathway, validation logic, and the core interaction abstractions of browsing, collections, and submission. The experimental layer comprises the iteration-specific interaction and mediation features introduced to test particular research questions, such as multi-criteria

filtering, LLM-assisted submission, LLM-mediated exploration via collection-bounded chat, and citation visibility manipulations. This distinction is important methodologically because it clarifies which parts of InterDev are treated as fixed experimental controls and which parts are intentionally varied across studies.

Item	Iteration 1	Iteration 2	Iteration 3
Source snapshot date	Sept 2023	Sept 2023	Sept 2023
3ie records ingested	10,431	10,431	10,431
AEA registry records ingested	2000	2000	2000
Total RDF triples	635,520	834,339	1,137,590
ERCT ontology version	ERCT V3	ERCT V4	ERCT V5
Triple store version	GraphDB 10.3	GraphDB 10.3	GraphDB 10.3
Backend environment	Flask 3.0 / Python 3.8	Flask 3.0 / Python 3.9	Flask 3.0 / Python 3.11
Frontend environment	React 18.2	React 18.2	React 18.2
Deployment mode	Local frontend / Institutional VM for backend	Institutional frontend / Institutional VM for backend	Institutional frontend / Institutional VM for backend

Table 6: Data and Implementation details across InterDev iterations

4.2 System Architecture and Technology Stack

This section describes the system architecture and technology stack underpinning the InterDev platform (see Figure 5). These components remain stable across all evaluation iterations and provide the infrastructural basis for the user-facing interaction layers examined in later chapters. As core architectural components and the underlying evidence base are held stable across iterations, observed differences in usability, workload, and confidence can be more plausibly attributed to interface-level and interaction-level changes rather than changes in the underlying data model or infrastructure. Code used for this implementation and to support the various evaluations (Chapter 5) is available at: <https://github.com/mmurtaghw/InterDevResearch>. Unless otherwise stated, the versions of each library mentioned here are consistent throughout each iteration of this thesis.

To make explicit the distinction between the stable platform and the experimental manipulations evaluated in Chapter 5, Table 7 separates the core infrastructure held constant across iterations from the interaction features varied by iteration.

Layer	Stable across iterations?	Components / examples	Role in the evaluation
Core platform	Yes	ERCT-aligned RDF data model; GraphDB triple store; Flask backend; backend-mediated SPARQL generation; React client structure; RDFLib validation; baseline evidence corpus (3ie, AEA, World Bank uplifted data); collection persistence model	Provides a constant architectural and data baseline so observed differences can be attributed to interface and workflow changes rather than backend or corpus changes
Iteration 1 interaction layer	No	Baseline evidence browsing; single-collection curation; manual submission; TTL export	Establishes baseline usability and develops the initial user requirements further through user observation.
Iteration 2 interaction layer	No	Multi-criteria filtering; multi-collection management; expanded export formats; LLM-assisted submission	Tests whether submission automation and interaction refinements improve efficiency and completion
Iteration 3 interaction layer	No	Collection-bounded chat; citation/provenance visibility manipulation; richer evidence comparison and synthesis support	Tests the effects of LLM mediation and provenance visibility on workload, confidence, calibration, and verification behaviour

Table 7: Stable platform infrastructure and iteration specific experimental interaction layers

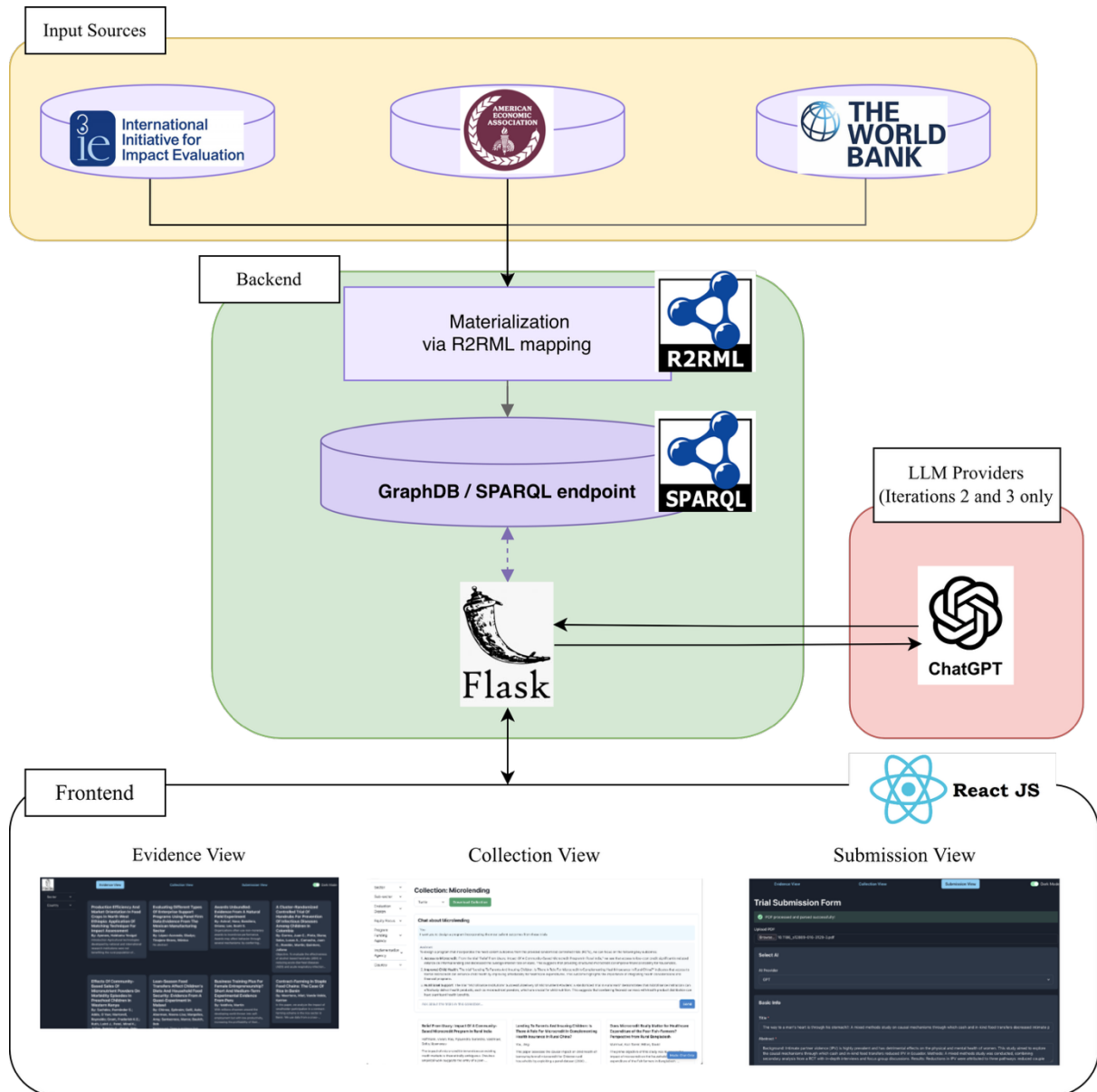


Figure 5: Overview of Framework

4.2.1 Client-Server Architecture

InterDev is implemented as a client–server web application. The frontend is developed using React 18.2 and is responsible for rendering user-facing views and handling interaction events. All data access and processing logic is handled by a Flask 3.0 backend service. Flask was chosen as it allows all backend logic to be written in Python, allowing access to a rich array of data focused libraries and tools supported by the language, while also being significantly more lightweight than comparable Python frameworks such as Django.

The backend acts as the sole mediator between the client and the knowledge graph. The frontend does not issue direct queries to the database. Instead, user interactions such as filtering, browsing, collection building, and submission are translated by the backend into structured SPARQL queries. This design enforces consistent query semantics, enables centralized validation and logging, and ensures that all experimental conditions operate over identical retrieval logic. This backend-mediated SPARQL design also improves experimental control. Because all user actions are translated through the same server-side query generation pathway, rather than allowing condition-specific or client-specific query behaviour, differences observed across iterations or interface conditions can be attributed more credibly to changes in interaction design, mediation, or provenance visibility rather than to differences in how data are retrieved. Centralising SPARQL generation also standardises filtering semantics, result ordering, validation, and logging, which makes the experimental conditions more directly comparable.

SPARQL query generation and normalization are handled within the backend Flask layer. This approach allows the system to evolve its interaction mechanisms across iterations while preserving a stable and reproducible data access pathway.

4.2.2 Knowledge Graph Backend

At the core of InterDev is a knowledge graph that represents impact evaluations and related contextual information in international development. The knowledge graph is stored in a GraphDB⁴ 10.3 triple store and serves as the authoritative source of evidence for all system functionality. GraphDB was selected over alternative RDF stores because it provided a stable, standards-compliant SPARQL endpoint, straightforward support for named-graph style organisation and update operations, and reliable local deployment for a research prototype whose priority was reproducible retrieval behaviour rather than store-level experimentation. In other words, the choice was pragmatic: the thesis does not evaluate triple-store performance as a research question, so the backend needed to be mature, interoperable, and sufficiently robust to support stable querying, ingestion, and collection management across iterations.

⁴ <https://graphdb.ontotext.com/>

All data within the system is represented using the Resource Description Framework (RDF), which functions as the canonical data representation. The structure and semantics of the graph are defined by the ERCT ontology⁵ discussed in Chapter 2 (Murtagh-White, 2021), which provides a shared schema for representing evaluations, interventions, outcomes, and contextual attributes. Using a common ontology enables integration of heterogeneous data sources while preserving semantic consistency across the graph.

The choice of RDF and a knowledge graph architecture supports interoperability, provenance tracking, and structured querying, properties particularly important in the context of evidence-based research. Maintaining a graph-based backend also allows curated subsets of evidence to be treated as explicit subgraphs, a concept that becomes central in later evaluation iterations in Chapter 5.

4.2.3 RDF Tooling and Validation

RDFLib⁶ 6.3.2, a Python library, is used throughout the InterDev platform to support RDF processing tasks. During data ingestion, RDFLib is employed to transform source data into ERCT-aligned RDF representations and to serialize these representations for storage in GraphDB. RDFLib is also used to parse and validate RDF data prior to insertion, testing for syntactic correctness and conformity with RDF standards. Mapping execution was performed using the R2RML engine Morph-KGC⁷, which materialises RDF triples directly from the mapping specifications

The same tooling is reused in later stages of the system to validate generated RDF outputs. While the mechanisms for producing these outputs vary across evaluation iterations, the validation step remains consistent. This reuse of RDFLib provides a uniform validation layer across both manual and automated workflows. Detailed discussion of LLM-assisted RDF generation and validation is deferred to later chapters, where these mechanisms are evaluated empirically.

⁵ <https://github.com/mmurtaghw/InterDevResearch>

⁶ <https://github.com/RDFLib/rdfliib>

⁷ <https://morph-kgc.readthedocs.io/en/stable/>

4.3 Data Sources and Knowledge Graph Construction

This section describes the data sources used to populate the InterDev knowledge graph and the process by which heterogeneous evaluation data is transformed into a unified, ontology-aligned representation. These data sources and construction mechanisms (see Figure 6) remain constant across all evaluation iterations so that changes observed in later studies arise from interaction and mediation design rather than differences in the underlying evidence base.

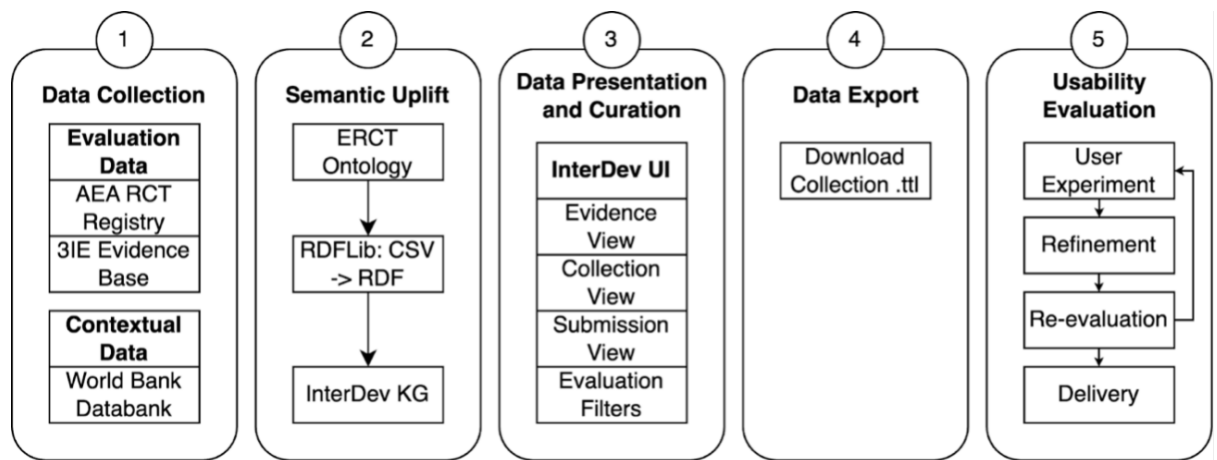


Figure 6: Overview of Implementation Methodology

4.3.1 Source Datasets

The InterDev knowledge graph integrates evaluation metadata and contextual indicators from multiple authoritative sources in international development research. The choice of data sources is guided by the requirements in Chapter 3 and by the need to hold the underlying evidence base constant across evaluation iterations. The 3ie evidence portal and the AEA RCT Registry are widely used, public-facing infrastructures for impact evaluation discovery and registration, and together provide complementary coverage of the evidence lifecycle described in Section 2.1.1: the AEA registry captures prospective trial records and design intent, while 3ie curates published evaluations and synthesis-facing metadata. Using both sources therefore exposes the fragmentation and heterogeneity that motivate InterDev (Chapter 1; Chapter 3), while still providing sufficiently structured fields to support ERCT-aligned uplift and faceted filtering (R1) without relying on manual extraction for the

baseline corpus. World Bank indicators are included to operationalise “contextualisation” as a queryable dimension (R1/R3), providing standardised country-level covariates that systematic reviews commonly use when stratifying or interpreting effects.

For example, a 3ie evidence record may provide the published evaluation title, sector, country, intervention category, outcome area, publication year, and links to reports or evidence products. An AEA RCT Registry record may instead provide prospective design information such as the trial title, intervention description, registration date, planned outcomes, experimental design, sample size, trial status, and implementing researchers. World Bank indicators add contextual variables that are not usually contained in evaluation records themselves, such as GDP per capita, poverty headcount, school enrolment, mortality, employment, population, or country income classification. In the InterDev knowledge graph, these fields are aligned to ERCT entities and properties so that a user can, for instance, filter evaluations by country and sector, inspect the associated intervention and outcome metadata, and compare studies against shared country-level development indicators.

Impact evaluation data were collected from the International Initiative for Impact Evaluation (3ie) evidence portal and the American Economic Association (AEA) Randomized Controlled Trial Registry. The 3ie dataset was obtained through automated scraping of the public evidence portal in January 2024 using Python libraries Selenium⁸ 4.12.0 and BeautifulSoup4⁹ 4.12.2, capturing structured metadata describing published evaluations. The AEA Registry data was acquired through downloadable CSV files, which provide detailed records of registered impact evaluations, including information on study design, interventions, and outcomes¹⁰.

To support contextualisation of evaluation evidence, additional data was sourced from the World Bank databank in early 2024. These indicators provide country-level contextual variables relevant to international development research, such as economic, demographic, and social measures.

⁸ <https://pypi.org/project/selenium/>

⁹ <https://pypi.org/project/beautifulsoup4/>

¹⁰ <https://dataverse.harvard.edu/dataverse/aearegistry>

4.3.2 Semantic Uplift and Ontology Alignment

All source datasets are transformed into a unified semantic representation through a semantic uplift process. This process maps raw source records into RDF triples that conform to the ERCT ontology. The uplift ensures that data originating from different repositories adheres to a shared schema and can be queried uniformly within the knowledge graph.

Initial uplifts used RDFLib to construct RDF representations from structured source formats such as CSV and scraped metadata. After initial prototyping, R2RML mappings are used to instantiate ontology-aligned resources for impact evaluations, interventions, outcomes, findings, provenance entities, and contextual country indicators. RDFLib continued to be used as supporting infrastructure for validation, graph-level post-processing, and serialization tasks. During uplift, entities such as impact evaluations, interventions, outcomes, and contextual attributes are instantiated as ontology-aligned resources, and relationships between these entities are explicitly encoded. Contextual World Bank indicators are integrated through the same pipeline: indicator data is fetched via API, normalised, and then uplifted via mapping rules into ERCT-compatible Indicator resources linked to country entities with provenance-preserving structure. All source datasets are transformed into a unified semantic representation using this R2RML-based semantic uplift process using declarative R2RML mappings. The scripts and mappings used for this are available in the thesis GitHub under each iteration folder¹¹.

4.3.3 Knowledge Graph Integration and Storage

Once uplifted, RDF data is inserted into the GraphDB triple store, which serves as the persistent backend for all system functionality. GraphDB supports scalable storage and SPARQL querying over the integrated knowledge graph, enabling retrieval behaviour across browsing, filtering, curation, and export operations for all iterations.

All evaluation iterations operate over the same knowledge graph structure and access pathways. Curated collections created by users are represented conceptually as subgraphs derived from the global ERCT graph, rather than as independent datasets. Collections are created and edited in the frontend, which calls backend endpoints that persist each

¹¹ <https://github.com/mmurtaghw/InterDevResearch>

collection as a KG resource (in this case a `skos:Collection`) and assert membership links to `erct:Trial` nodes in GraphDB. The backend handles create/add/remove operations as SPARQL updates and returns collection views by querying the same ERCT graph, so UI state is derived from the KG. This design allows subsets of evidence to be manipulated, exported, or analysed while maintaining explicit links to the original source data. This provides a consistent foundation for the experimental manipulation of interaction approaches examined in later chapters. With the evidence base constructed and normalised into ERCT-aligned RDF, the central remaining question becomes interactional: how users discover, bound, and reuse evidence without needing to understand RDF, OWL, or SPARQL. Section 4.4 therefore describes the stable interaction abstractions that realise R1–R3 at the user interface layer.

4.4 Core Interaction Abstractions

This section describes the core interaction abstractions that underpin all versions of the InterDev platform. These abstractions define the functional roles through which users engage with the knowledge graph, independent of specific interface layouts or interaction techniques. While individual affordances within each abstraction evolve across evaluation iterations, the abstractions themselves remain stable and provide a consistent conceptual model for browsing, curating, and contributing evidence.

4.4.1 Global Interface Structure

InterDev is organised around a global interface structure that separates user interaction into distinct functional modes. This structure supports three primary activities: evidence browsing, evidence curation, and evidence contribution. Each activity is accessed through a dedicated view, allowing users to shift between tasks without altering the underlying data representation. Navigation mechanisms support view switching while maintaining continuity of user context, such as selected evaluations or curated collections. This separation of concerns allows interaction designs to evolve independently within each view.

4.4.2 Evidence View

The Evidence View provides the primary mechanism for discovering and inspecting impact evaluations stored in the knowledge graph. Its core function is to enable users to explore the evidence base by applying constraints and examining individual impact evaluations.

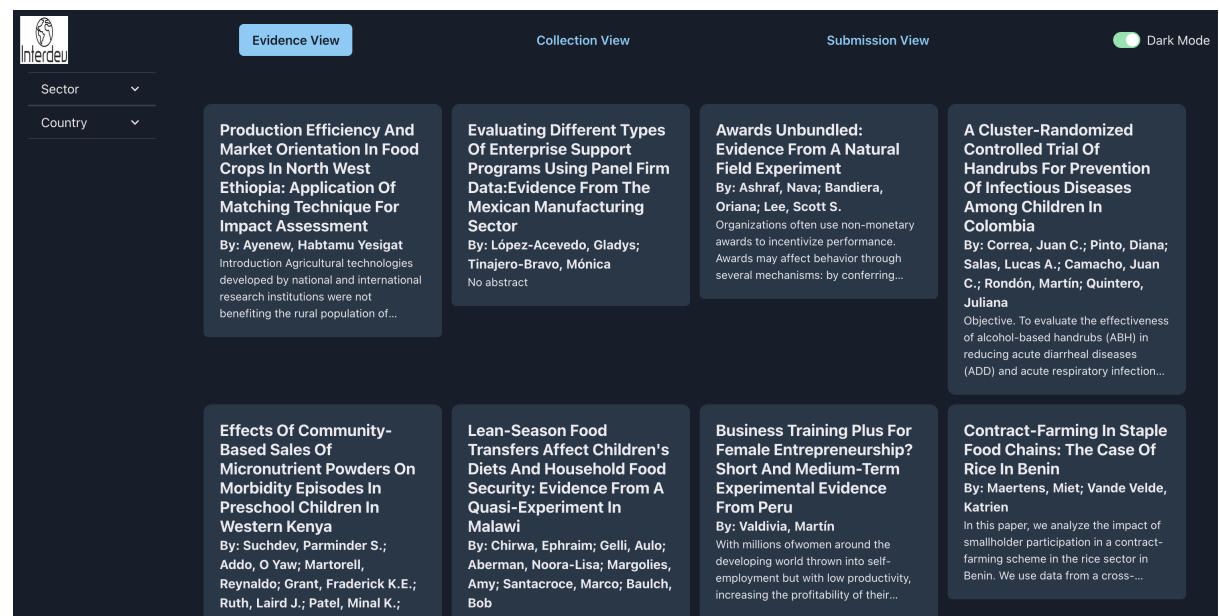


Figure 7: Evidence view of InterDev Version 1

Within this view, users can refine the set of visible evaluations by applying filter-based constraints derived from structured attributes in the knowledge graph. These constraints allow users to narrow the evidence space according to research-relevant criteria without requiring knowledge of the underlying ontology or query language.

The Evidence View also supports inspection of individual evaluations. Users can examine detailed metadata associated with a selected impact evaluation, including information about study design, interventions, outcomes, and contextual attributes. This inspection capability allows users to assess relevance and quality before incorporating evaluations into curated collections or analytical workflows. This view directly operationalises R1 by exposing ontology-aligned facets as user-facing filters, while maintaining stable semantics through backend-mediated query generation.

4.4.3 Collection View

The Collection View supports the curation of user-defined subsets of evaluations. A collection represents an explicit selection of impact evaluations drawn from the global knowledge graph and serves as a working set for comparison, synthesis, and export.

Conceptually, a collection corresponds to a subgraph of the ERCT knowledge graph. While users interact with collections through familiar interface mechanisms, the underlying abstraction preserves the structural relationships between selected evaluations and their associated metadata. This design enables curated evidence to be treated as a coherent unit while maintaining traceability to the original sources.

Collections form the basis for downstream activities such as cross-impact evaluation comparison, data export, and, in later iterations, assisted synthesis. By making collections an explicit abstraction, the platform supports exploration over selected evidence without modifying the global knowledge graph. Collections operationalise R2 and R3 by turning ad hoc analysis subsets into explicit, reusable units with persistent identifiers that remain across sessions.

4.4.4 Submission View

The Submission View provides the mechanism through which new evaluations are added to the InterDev knowledge graph. Its primary role is to support structured contribution while maintaining alignment with the ERCT ontology.

All submission workflows adhere to a human-in-the-loop principle. Users remain responsible for reviewing and confirming evaluation metadata before it is integrated into the knowledge graph, regardless of how that metadata is produced. This ensures that contributions remain auditable and consistent with the platform’s emphasis on transparency and provenance.

While the degree of LLM-assistance within the submission process varies across evaluation iterations, the Submission View consistently functions as the boundary between external evaluation materials and their structured representation within the knowledge graph.

4.5 Iteration-Specific Implementation Extensions

While the core architecture, data model, and interaction abstractions described above remain stable across all versions of InterDev, successive evaluation iterations introduce targeted implementation extensions to address usability issues identified in earlier studies and to support new research questions.

- **Iteration 1** implements baseline evidence browsing and filtering, single-collection curation, manual metadata submission, and RDF export, establishing the initial functional platform evaluated for usability breakdowns.
- **Iteration 2** extends the submission workflow with LLM-assisted submission, introduces enhanced collection management and expanded export formats, and supports controlled comparison between LLM-assisted and manual submission workflows.
- **Iteration 3** introduces an LLM-mediated exploration layer over curated collections, manipulates citation visibility as an interface-level condition, and adds support for evidence comparison and synthesis tasks under controlled experimental conditions.

Detailed descriptions of these extensions, along with their empirical evaluation, are provided in the corresponding sections in the next evaluation chapter.

Chapter 5

Evaluation

This chapter reports an iterative evaluation of InterDev that tracks how successive design interventions changed user performance and experience in evidence discovery, curation, submission, and performance on synthesis-oriented tasks. The evaluation is organised as a progression of three iterations that mirror the thesis’s core tension: improving usability and efficiency while preserving auditability and appropriate reliance. Iteration 1 establishes baseline usability and localises interaction breakdowns in the core faceted browsing, collection curation, and manual submission workflows. Iteration 2 evaluates whether an LLM-assisted, human-reviewed submission pipeline can reduce contribution effort and increase completion without sacrificing perceived quality. Iteration 3 shifts from usability and efficiency to epistemic outcomes, examining how LLM-mediated exploration with provenance cue visibility affect workload, verification behaviour, confidence calibration, and synthesis quality during evidence use. In addition to the three InterDev iterations, the chapter also reports a complementary cross-domain study in digital libraries, included as a companion evaluation of whether the underlying interaction modality shows signs of broader feasibility beyond international development.

Across the three InterDev iterations, the architectural baseline and evidence corpus are held stable (Chapter 4) so that changes in outcomes can be interpreted primarily as consequences of interaction and mediation design. At the same time, the measurement strategy intentionally broadens as the system matures. Across these iterations, the implementation of InterDev is refined in response to empirical findings, and each evaluation phase motivates the next set of design decisions and defines new requirements for the system to fulfil. In addition to reporting usability outcomes, the chapter examines how interaction modality, automation, and provenance cues influence user confidence, trust calibration, workload, and synthesis practices in evidence exploration.

Table 8 summarises the evaluation programme across InterDev’s three iterations and the complementary cross-domain study, indicating for each study the primary goal, design, participant sample, core measures, and principal outcome.

Iteration	Main focus	Study design / conditions	Participants	Core measures	Key outcome
Iteration 1	Baseline usability of evidence browsing, filtering, collection curation, and manual submission	Formative usability study; single baseline system; supervised think-aloud	10	Task completion time, PSSUQ, assistance/intervention counts, think-aloud, open-ended responses	Identified major usability issues in filtering, collection management, export, and manual submission
Iteration 2	Whether LLM-assisted submission reduces contribution effort relative to manual submission	Between-subjects comparison: manual submission vs LLM assisted submission	20	Task completion time, completion rate, PSSUQ, perceived accuracy, subjective trust, open-ended responses	LLM-assisted submission improved completion and reduced time-on-task relative to manual submission
Iteration 3	Epistemic effects of LLM-mediated exploration and provenance cue visibility	Within-subjects, counterbalanced design: C0 baseline browsing, C1 LLM-mediated exploration, C2 LLM-mediated exploration + citations	21	Task performance rubric, task time, NASA-TLX, Trust in Automation, task-level confidence, confidence calibration, PSSUQ, interviews	Chat reduced workload, but did not improve task performance; citations reduced confidence and worsened calibration without increasing trust
Cross-domain study	Feasibility of the interaction modality beyond international development	Within-subjects, counterbalanced design across four interfaces: keyword, semantic, conversational, conversational-graph	28	Task performance, task time, NASA-TLX, confidence, calibration, qualitative interview data	Conversational interaction improved exploratory support and lowered workload, but richer interfaces introduced trade-offs in confidence and cognitive effort

Table 8: Overview of InterDev Evaluation Iterations

5.1 Evaluation methodology for InterDev framework

As discussed in Section 1.5, the overall design and implementation of the InterDev system employs a six stage UCD methodology following established principles (Boren & Ramey, 2000; DIS, 2009; Kujala, 2003; J. R. Lewis, 2012; Nielsen, 1994; Norman, 1988; Preece et al., 2015; Saffer, 2010). This chapter focuses on how that framework is instantiated through the iterative evaluation rounds conducted during the later stages of the process.

While the evaluation programme is grounded in a User-Centred Design methodology, later iterations extend beyond traditional usability assessment. As core interaction breakdowns are resolved, the evaluation focus shifts toward understanding how LLM-mediated interaction, specifically LLM-assisted submission and LLM-mediated exploration, affect cognitive workload, confidence, trust calibration, and performance on synthesis-oriented tasks. Accordingly, the methodological framework evolves to incorporate controlled experimental designs and additional quantitative measures alongside standard usability instruments.

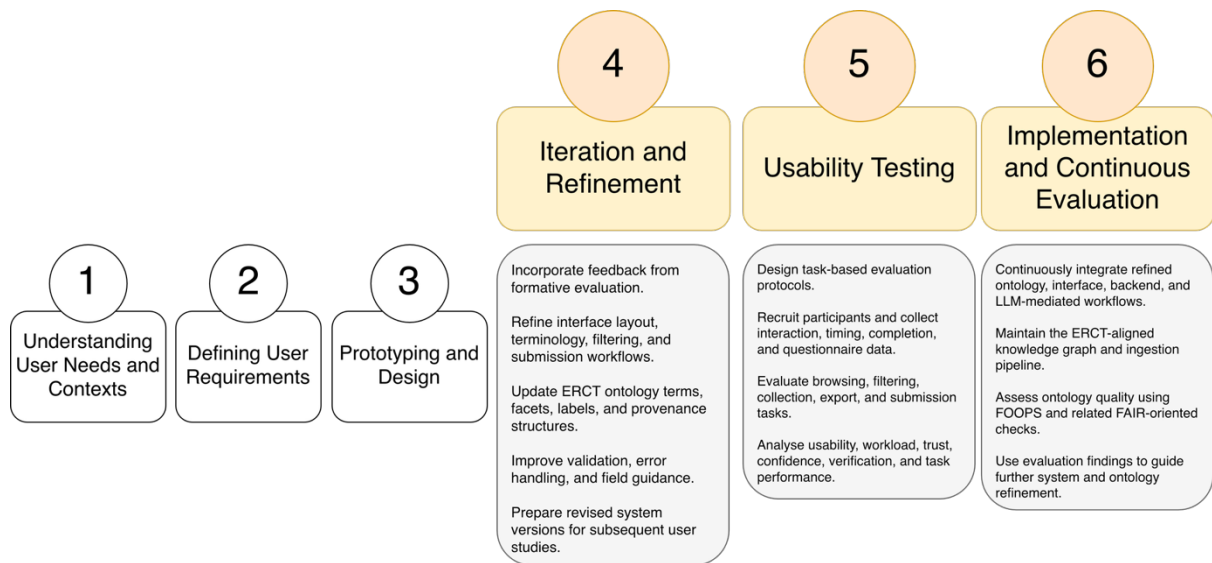


Figure 8: InterDev Methodological Process – Usability Testing, Implementation and Continuous Evaluation

The evaluation programme places particular emphasis on the Stage 5–6 cycle: usability testing followed by implementation and continuous evaluation, where findings from each round motivate subsequent design refinements. This loop is adapted from established usability evaluation frameworks (Boren & Ramey, 2000; J. R. Lewis, 1995, 2012). Figure

8 provides a focused view of this cycle within the broader six-stage framework introduced in Section 1.5.

As core interaction breakdowns are addressed, the evaluation focus expands beyond baseline usability to examine the effects of LLM-mediated interaction, specifically LLM-assisted submission and LLM-mediated exploration, on cognitive workload, confidence, trust calibration, verification behaviour, and performance on synthesis-oriented tasks. Accordingly, later studies incorporate controlled experimental manipulations and additional quantitative measures alongside core usability instruments; these are detailed in Sections 5.2–5.3.

5.2 Experimental Setup

5.2.1 Participant Setup

Participants invited to participate in the experiments were offered either online or in-person sessions, dependent on the iteration. If a participant opted to participate in the experiment remotely, a link was sent to the participant to access the InterDev platform remotely. The participant was invited to a Zoom call with the experimenter and asked to share their screen and audio. If the participant opted for an in-person session, they were invited to a private booked space at the ADAPT centre at Trinity College Dublin, where they were invited to use InterDev on the experimenters computer. This hybrid setup allowed for a wider range of participants to partake in the experiment, allowing participants to join regardless of location.

The remaining experimental conditions were the same for both online and in-person participants. If not already completed, the participant was asked to sign the consent form at the onset of the experiment. The participants were then asked to complete a series of tasks, which were selected to reflect the workflow of researchers engaging with the framework, ranging from filtering evaluations on the graph, creating collections and contributing new evaluation data to the graph. Participants were briefed on the steps involved in the experiment and the main evaluation instruments, the think-aloud protocol and the PSSUQ. Participants were notified that the experiment would be recorded for the purpose of gathering data from the think-aloud protocol and downstream thematic analysis. The first

evaluation used Google Forms¹² to structure the evaluation and record results. Later evaluations (from iteration 2 onwards) used Qualtrics¹³ to structure the evaluation and record the results, as this allowed for features such as conditional workflows for Latin square condition assignments and the automated recording of task time.

5.2.2 Participant Recruitment

The study involved recruiting participants with a background in social science research, ensuring foundational understanding of research methodologies and theoretical frameworks. Preference was given to individuals with experience in the field of international development, as their expertise and familiarity with the domain were deemed most suited for evaluating the InterDev platform's usability within its intended context.

Participants were invited via email to partake in the usability study. Invitations outlined the purpose of the study, the expected duration, and the hybrid nature of the experimental setup (allowing for both in-person and remote participation, where applicable). The recruitment email emphasised the importance of domain-specific insights in enhancing the framework's alignment with user needs.

To ensure a diverse and representative sample of the target user base, the recruitment strategy included outreach to professional networks, academic institutions, and organisations specialising in international development. This approach facilitated the inclusion of participants with varying levels of familiarity with digital tools and platforms, thereby capturing a broad spectrum of user experiences. Different participants were used for each round of evaluation.

5.3 Evaluation Instruments

The evaluation programme uses a set of core instruments applied across all iterations, supplemented by iteration-specific measures introduced as the research focus evolved.

¹² <https://docs.google.com/forms/>

¹³ <https://www.qualtrics.com>

5.3.1 Post Study System Usability Questionnaire

The Post-Study System Usability Questionnaire (PSSUQ) is used as a quantitative instrument to measure participants' perceived usability and satisfaction with the InterDev platform following task completion. Developed by Lewis (J. R. Lewis, 1992, 1995, 2002), the PSSUQ is a validated tool widely used in usability evaluation and assesses user experience across three dimensions: System Usefulness, Information Quality, and Interface Quality. Items are rated on a Likert scale, enabling both overall usability scoring and subscale analysis.

This thesis uses PSSUQ v2, which comprises 19 items: System Usefulness (Q1–Q8), Information Quality (Q9–Q15), and Interface Quality (Q16–Q19), alongside an overall score computed across Q1–Q19 (J. R. Lewis, 2002). PSSUQ responses provide a consistent basis for comparing perceived usability across iterations and experimental conditions, and are interpreted alongside behavioural measures (e.g., task time) and qualitative feedback.

PSSUQ is not without limitations. As a post-task self-report instrument, it may not capture real-time usability challenges or the cognitive processes users experience during task execution, and it can be affected by recall effects and social desirability bias (Hodrien & Fernando, 2021). For this reason, PSSUQ is used in conjunction with complementary qualitative methods and behavioural measures throughout the evaluation programme.

5.3.2 Evaluation Metrics

The evaluation of InterDev employed a mixed set of quantitative and qualitative measures. A small set of core measures was retained across all three iterations to preserve comparability, while additional measures were introduced in later iterations to reflect the shift from baseline usability assessment to LLM-assisted submission and, subsequently, to the epistemic effects of LLM-mediated interaction.

5.3.2.1 Quantitative Measures

Task Completion Times Task completion time was captured for each evaluation round as a measure of interaction efficiency. The logging method varied by study design (observer

timing in early studies; automated logging or survey timestamps in later studies), but the construct remained consistent: time required to complete defined workflows in InterDev.

Post-Study System Usability Questionnaire (PSSUQ) Participants complete the 19-item PSSUQ after task completion (see Section 5.3.1 for instrument details and scoring). PSSUQ scores are analysed at the item level and at the standard subscale level (System Usefulness, Information Quality, Interface Quality), alongside an overall usability score.

5.3.2.2 Qualitative Measures

Qualitative feedback was collected to contextualise quantitative outcomes and identify interaction breakdowns not captured by questionnaires alone. This included open-ended responses associated with post-task instruments, and, where employed, verbal protocols or interviews. The specific qualitative modality varied by iteration (e.g., think-aloud in early formative testing versus structured interviews in later evaluative rounds) and is detailed in the iteration-specific methodology sections.

5.3.2.3 Supplementary measures introduced by iteration

As the evaluation focus evolved, additional measures were incorporated to address iteration-specific research questions. The full breakdown of each measure and instrument used across iterations is given in Table 9.

- **Formative usability diagnostics (Iteration 1).** Assistance/intervention counts and concurrent verbalisation were used to localise interface breakdowns and guide redesign priorities.
- **LLM-assisted submission perceptions (Iteration 2).** Measures were added to capture perceived accuracy of extracted submission metadata and trust in the submission workflow under manual versus LLM-assisted submission conditions.
- **Epistemic and workload measures (Iteration 3).** Subjective workload was measured using NASA–TLX; trust in automation was measured using the Trust in Automation scale; confidence ratings were collected per task and analysed using calibration metrics (Brier score). Task responses were also scored using a structured rubric to capture evidence use and synthesis quality.
- **Cross-domain measures (digital-library companion study).** The cross-domain evaluation added the same higher-order measures used in the later InterDev

iterations: NASA–TLX workload, self-reported confidence, confidence calibration (Brier score), task completion time, task performance scoring (constraint checks and rubric-based scoring for synthesis/verification tasks), and trust/transparency Likert items, supported by post-condition reflections and interviews.

Measure / Instrument	InterDev Iteration 1	InterDev Iteration 2	InterDev Iteration 3	Cross- domain study
Task completion time	✓	✓	✓	✓
Task performance / rubric scoring	—	—	✓	✓
PSSUQ	✓	✓	✓	—
Open-ended questionnaire responses	✓	✓	✓	✓
Think-aloud protocol	✓	—	—	—
Assistance / intervention counts	✓	—	—	—
Perceived accuracy rating	—	✓	—	—
Subjective Trust rating	—	✓	—	—
NASA–TLX workload	—	—	✓	✓
Trust in Automation scale (Jian et al.)	—	—	✓	—
Task-level confidence ratings	—	—	✓	✓
Confidence calibration (Brier score)	—	—	✓	✓
Controlled experimental manipulation (multiple conditions)	—	✓	✓	✓
Qualitative interview data	—	—	✓	✓
Thematic analysis	✓	—	✓	✓
Mixed-effects / regression modelling	—	—	✓	✓

Table 9: Evaluation measures and instruments used throughout evaluation iterations and the cross-domain study

5.3.3 Data Analysis

Across all evaluation rounds, analysis followed a consistent mixed-methods strategy that integrates quantitative usability/efficiency outcomes with qualitative evidence about user experience and interaction breakdowns. Quantitative measures provided comparable

benchmarks across iterations (e.g., task completion time and PSSUQ), while qualitative data provided explanatory depth regarding why breakdowns occurred and how participants interpreted system behaviour. Although the specific instruments varied by iteration (Table 9), the analytic approach was intentionally held constant: (i) compute and compare quantitative outcomes and (ii) apply a common qualitative coding and theme-development procedure to interpret user feedback and observed behaviour.

5.3.3.1 Quantitative analysis

Quantitative data were analysed in two layers. First, all studies report descriptive summaries (distributions, central tendency, and variability) for core measures such as task completion time and questionnaire responses. Second, where the study design included controlled experimental manipulation (Iterations 2 and 3, and the cross-domain study), inferential comparisons were performed to estimate condition effects. In later evaluations with repeated-measures designs (Iteration 3 and the cross-domain study), mixed-effects regression modelling was used to account for within-participant dependence and to control for task-level difficulty where appropriate. These modelling choices are reported within the corresponding iteration sections to maintain transparency and avoid overstating analytic complexity in earlier formative rounds.

5.3.3.2 Qualitative analysis

Qualitative data collected across the evaluation programme, including open-ended questionnaire responses, concurrent verbal protocols (where employed), and post-study interviews (where employed), this research employs thematic analysis as outlined by Braun and Clarke ((Braun & Clarke, 2006; Clarke et al., 2015)) and refined by Nowell et al. (Nowell et al., 2017). This methodology is well-established in usability studies and ensures a rigorous and reliable analysis of the data (Rosen et al., 2023). Below, the six-step process for thematic analysis is described in detail to set the stage for the evaluation. The same six-step process was applied across evaluations to ensure consistency and traceability of interpretation, regardless of the specific qualitative data source used in a given iteration.

1. Familiarisation with the Data

Qualitative responses were collated for each study round and reviewed repeatedly to build an initial understanding of recurring issues, successful interaction patterns, and points of

friction. Where audio data were collected, recordings were transcribed and checked for accuracy before coding began. Preliminary notes and observations were documented, highlighting recurring thoughts, challenges, and user suggestions. These notes formed the initial context for the coding phase.

2. Generating Initial Codes

Data segments were assigned descriptive codes capturing usability breakdowns, interpretations of system behaviour, and participant suggestions for improvement. NVivo was used to improve traceability and ensure consistent coding across datasets. For example, comments like “I can’t find where to filter by country” were coded as “Filtering Challenges,” while “This looks neat and organised” was tagged as “Positive Interface Feedback.” This coding framework helped break down user feedback into actionable units, which were later grouped into broader themes.

3. Searching for Themes

Codes were grouped into candidate themes that reflected recurring patterns in user experience, such as navigation friction, feature discoverability, provenance-related concerns, and synthesis support. Candidate themes were refined iteratively as additional data were coded. The tagged codes were grouped into broader themes that reflected recurring patterns in the data. For InterDev, themes such as “Navigational Challenges,” “Visual Design Issues,” “Positive Task Flow Experiences,” and “Feature Discoverability” emerged. These themes encapsulated the core usability issues and strengths of the platform, aligning user feedback with areas for improvement.

4. Reviewing Themes

Themes were reviewed against the source data to confirm internal coherence and adequate evidential support. Themes that were weakly supported were merged, split, or removed to improve clarity and specificity.

5. Defining and Naming Themes

Themes were defined with clear scope boundaries and supported with representative excerpts. Definitions were written to align with the focus of each evaluation round, while

maintaining a stable interpretation of theme labels across the broader programme. Each theme was defined in detail and given a clear, descriptive name. For instance, the theme “Guidance and Navigation” was defined as “Comments on the clarity of instructions, ease of navigation, and suggestions for improving user guidance, such as better task prompts and visual cues.” Definitions included examples from the data, such as direct quotes or observed behaviours, to illustrate the scope of each theme.

6. Reporting findings and integrating with quantitative results.

Themes were synthesised alongside quantitative outcomes to produce a holistic interpretation of system performance and user experience. Where appropriate, code frequencies and theme prevalence were summarised using scripts to support transparent reporting. Findings were then translated into iteration-specific design recommendations.

5.4 Usability Evaluation Iteration 1

This section describes the initial usability evaluation. This work was peer-reviewed, presented and published at the 2024 Visualization and Interaction for Ontologies, Linked Data and Knowledge Graphs (VOILA) workshop at the International Semantic Web Conference (ISWC) (Murtagh-White et al., 2024). Ethics for this study was approved by the Research Ethics Committee at Trinity College Dublin, REF2847.

5.4.1 System Under Evaluation

The evaluation was conducted on the baseline InterDev system described in Chapter 4 (Sections 4.2–4.5), which provides evidence browsing, collection curation, manual submission, and TTL export over an ERCT-aligned RDF knowledge graph.

5.4.2 Experimental Design

The study used a mixed-methods design. Ten participants, comprising PhD researchers and social science researchers with no prior semantic web experience, completed ten tasks using the InterDev platform. The aim was to evaluate whether users without specialist semantic web knowledge could effectively navigate, interact with, and use the platform for research tasks. Task completion times were recorded by the observer as a measure of efficiency.

Table 10 displays the task list and corresponding requirement, identified as per section 3.2. The tasks involved in the evaluation were as follows: selecting “Evidence View” from the navigation bar and waiting for the information to appear (T1), selecting any impact evaluation from the evidence view and viewing its associated information (T2), noting the sector of the selected impact evaluation (T3), filtering the impact evaluations in the evidence view by the noted sector until only impact evaluations from that sector appear (T4), adding four impact evaluations from this selection to the collection and confirming their presence in the “Collection View” (T5), returning to the evidence view and filtering for both a country and a sector, adding at most four more impact evaluations to the collection, and confirming their presence in the “Collection View” (T6), going to the “Collection View,” filtering the collection by any property, and downloading the collection (T7), submitting a new impact evaluation with any data in the “Impact Evaluation Submission” view (T8), finding the submitted evaluation data in the “Evidence View” (T9), and finally, downloading the evaluation data in the .ttl format (T10).

As shown in Table 9, the measures used in the user evaluation were: PSSUQ, task completion time, assistance/intervention counts, open-ended questionnaire responses. The evaluation also included a Think-Aloud protocol (Boren & Ramey, 2000). During the evaluation tasks, the participants verbalized their thoughts and actions, providing real-time feedback on their experiences and any difficulties encountered. These were transcribed and systematically analysed using the process outlined in Section 5.3.3.2. The transcriptions were used to identify recurring themes such as frustrations, successes, or suggestions for improvement, offering insights into the user experience. The think-aloud protocol was implemented purposefully in this first iteration in order to gather rich, actionable user data on the fundamental design of the system that could inform the overall design philosophy of subsequent iterations and evaluations.

Task Number	Task Description	Evaluated Requirement
T1	Select Evidence View from the navigation bar	Filtering and Context
T2	Select an impact evaluation from the evidence view	Filtering and Context
T3	Identify the sector of the selected impact evaluation	Filtering and Context
T4	Filter impact evaluations by the identified sector	Filtering and Context
T5	Add four impact evaluations to the collection	Granular and Aggregated Data
T6	Filter impact evaluations by country and sector, and add up to four more to the collection	Granular and Aggregated Data
T7	In Collection View, filter the collection by any property and download the collection	Granular and Aggregated Data
T8	Submit a new impact evaluation in Submission View	Cross-Study Comparison
T9	Find the submitted impact evaluation in Evidence View	Cross-Study Comparison
T10	Download the evaluation data in .ttl format	Granular and Aggregated Data

Table 10: Task list and associated requirements.

Instances in which participants encountered issues and required assistance were also recorded for each task. After completing the tasks, participants completed the PSSUQ. Together, these measures provided a mixed-methods assessment of platform usability and informed subsequent redesign decisions. The data produced as part of this evaluation is available on the thesis GitHub under the Iteration 1 folder¹⁴.

¹⁴ <https://github.com/mmurtaghw/InterDevResearch>

5.4.3 Quantitative Results

Figure 9 illustrates the box plot of time spent to complete each task. Tasks such as selecting the “Evidence View” from the navigation bar (Task 1), selecting any impact evaluation from the evidence view (Task 2), and noting the sector of the impact evaluation (Task 3) have low median completion times and minimal variability, indicating that users found these tasks straightforward and easy to complete. However, tasks involving filtering and managing collections presented more challenges. For instance, Task 4, which requires filtering impact evaluations for the noted sector, shows moderate median completion time with some variability, suggesting users found the filtering function somewhat challenging. Task 5, which involves adding four impact evaluations to the collection, and Task 6, which includes filtering for both a country and a sector, both exhibit higher median completion times and significant variability, indicating these tasks were particularly difficult for users. Other tasks, such as submitting a new impact evaluations (Task 8) and finding the submitted evaluation in the evidence view (Task 9), also show higher median completion times and some outliers, reflecting challenges in the submission process and locating submitted

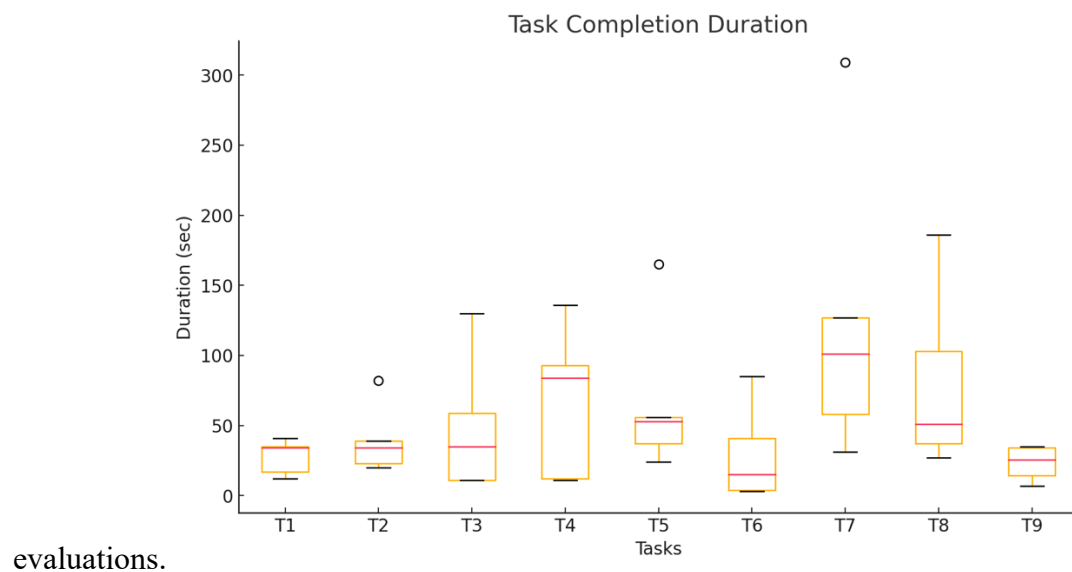


Figure 9: Boxplot of time spent to complete each task in the usability evaluation.

Figure 10 shows the intervention count for each task, providing further insights into task difficulty. Task 7, which involves filtering the collection by any property and downloading it, had the highest number of interventions, suggesting it was particularly challenging for users. Tasks 2, 3, and 5 had moderate intervention counts, indicating these tasks presented some challenges but were generally manageable. Tasks 1, 4, and 8 had lower intervention

counts, suggesting these tasks were relatively straightforward for users. Tasks 6, 9, and 10 had no recorded interventions, indicating that these tasks were the easiest for users to complete independently.

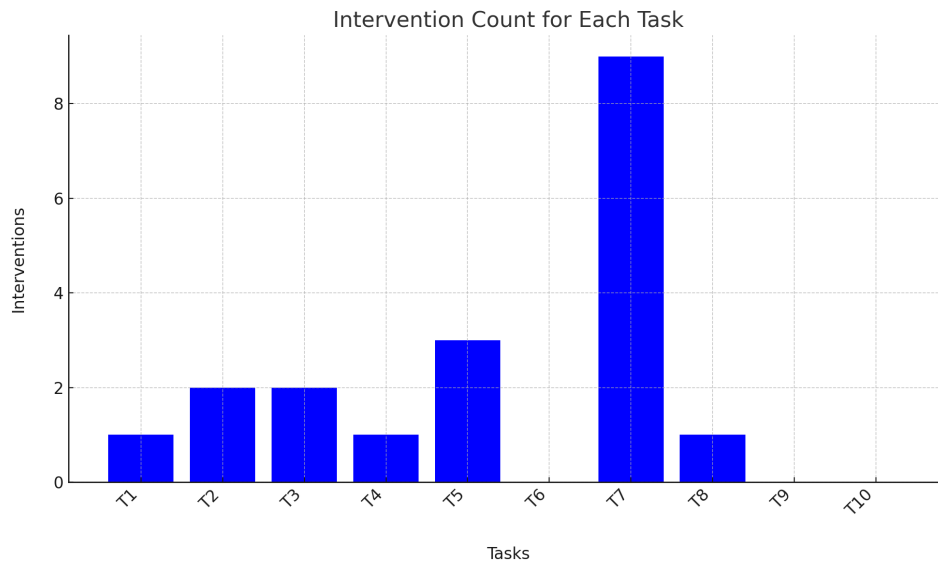


Figure 10: Bar chart of the number of interventions required for each task.

The analysis of the PSSUQ data seen in Figure 11 indicates that users generally find the system functional, with lower scores reflecting better usability and satisfaction. In particular the aggregate based scores related System Usefulness (SysUse), information Quality (InfoQual) and Interface Quality (IntQual) and Overall (as seen on right hand side of the Figure) are noteworthy. However, significant variability in satisfaction levels related to specific questions were observed. Notably, questions related to error messages (Q9) and overall satisfaction (Q19) exhibit higher scores and outliers, suggesting inconsistent user experiences in these areas. This inconsistency underscores the need for targeted improvements in error message clarity and overall system responsiveness.

The implications of these findings suggest that while the InterDev platform serves its primary purpose, there is substantial room for improvement. Enhancing error message clarity can significantly reduce user frustration and improve task efficiency, allowing more intuitive interaction with the system.

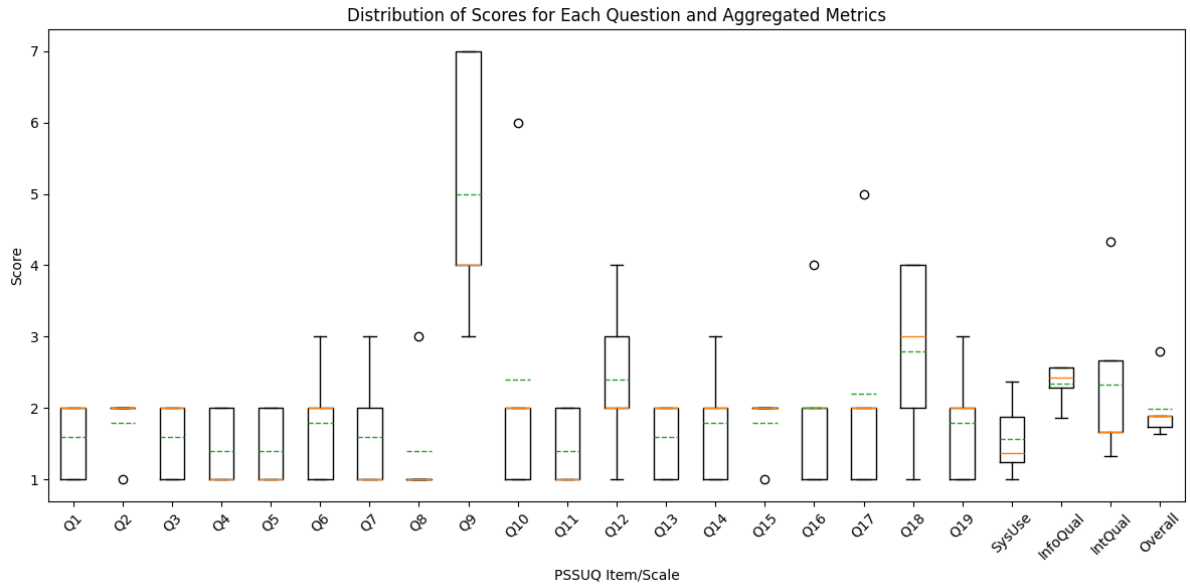


Figure 11: Boxplot of PSSUQ scores, based on a Likert 7-point scale where lower values indicate a higher satisfaction. System Usefulness (SysUse), information Quality (InfoQual) and Interface Quality (IntQual) and Overall are aggregated metrics based on questions 1-8, 9-15, 16-18 and 1-19 respectively.

5.4.4 Qualitative Results

Table 11 summarizes the thematic analysis for the first iteration of InterDev user testing, providing further insights into user feedback. Usability and Interface Design (UID), which encompasses overall design, intuitiveness, and ease of use, had the highest frequency with 19 mentions, indicating that users frequently commented on the visual layout, ease of finding information, and general user experience. Guidance and Navigation (GN) had 13 mentions, highlighting user comments on the clarity of instructions, ease of navigation, and suggestions for improving user guidance, such as better task prompts and visual cues. Functionality and Features (FF) was mentioned 12 times, reflecting feedback related to the platform’s functionalities, including search capabilities, filtering options, and specific features like collection management and submission forms. Efficiency and Performance (EP), with 9 mentions, included observations related to the speed and efficiency of completing tasks, as well as any technical issues or bugs encountered during use.

Theme	Code Description	Frequency
Usability and Interface Design (UID)	Overall design, intuitiveness, and ease of use of the platform's interface. This includes feedback on visual layout, ease of finding information, and general user experience.	19
Guidance and Navigation (GN)	Comments on the clarity of instructions, ease of navigation, and suggestions for improving user guidance, such as better task prompts and visual cues.	13
Functionality and Features (FF)	Feedback related to the platform's functionalities, such as search capabilities, filtering options, and specific features like collection management and submission forms.	12
Efficiency and Performance (EP)	Observations related to the speed and efficiency of completing tasks, as well as any technical issues or bugs encountered during use.	9

Table 11: Thematic analysis summary for the first iteration of InterDev user testing.

The initial evaluation of InterDev demonstrates its potential to enhance data discovery and usability for researchers in international development. While users found the platform generally functional, significant improvements are needed, particularly in filtering, error message clarity, search capabilities, and overall interface design. Quantitative and qualitative feedback from our user study highlighted key areas for enhancement, such as better guidance, improved navigation, and more intuitive features. These insights will guide the iterative refinement of InterDev to better meet user needs. While InterDev shows promise, further continuous user-centred development is required. Future iterations will address identified challenges, further refine user needs, and aim to improve the user experience and maximize the platform's utility in making international development data more accessible and actionable.

5.4.5 Limitations

This first evaluation was intentionally formative, aiming to establish baseline usability, identify issues, and generate redesign targets rather than to support strong generalisable or causal claims. The small sample size ($n = 10$) limits statistical power and the stability of quantitative summaries such as task times, intervention counts, and PSSUQ scores, and the participant pool may not fully represent the diversity of the intended user base. The findings

therefore primarily indicate likely high-friction workflows rather than population-level usability estimates.

The supervised, think-aloud study format also constrains validity. Think-aloud can slow task execution and change exploration strategies, inflating task times relative to naturalistic use, while the presence of an experimenter and the availability of assistance may reduce the real-world consequences of user issues. As a result, both efficiency measures and intervention counts should be interpreted as indicative rather than definitive.

The task set focused on foundational workflows such as browsing, filtering, collection management, export, and manual submission, and does not capture the full complexity of end-to-end evidence synthesis, such as extended comparison, verification, and integration of findings. In addition, measurement choices impose limits: observer-timed task durations and assistance counts are coarse proxies for difficulty, and PSSUQ is a retrospective self-report that cannot localise specific sources of friction. Finally, because the study did not include controlled interface conditions or experimental manipulations, the quantitative results are descriptive rather than causal, motivating Iteration 2's shift toward controlled comparisons and automated logging.

5.4.6 Conclusions

The first evaluation established a baseline picture of InterDev's usability and identified the main points of friction in the initial platform. Empirically, participants were generally able to complete core browsing and inspection tasks with relatively low difficulty, indicating that the basic evidence-view model was understandable without prior semantic web expertise. In contrast, filtering, collection management, and manual submission tasks showed longer completion times, greater variability, and higher intervention needs, indicating that these workflows imposed substantially greater interaction cost. Questionnaire responses and thematic analysis converged on similar problem areas, particularly around filtering clarity, navigation support, feature discoverability, error communication, and the burden of manual data entry.

These findings suggest two main design implications. First, discovery and curation workflows required clearer interaction support, particularly for filtering, collection

management, and system feedback. Second, the manual submission workflow emerged as a deeper bottleneck. Performance on Tasks 8 and 9 suggests that users were not simply hindered by interface design, but by the difficulty of translating evaluation information into the structured metadata expected by the platform. This implies that submission improvements would require not only interface refinement but also stronger structuring and assistance mechanisms.

Accordingly, Iteration 2 focused on two response strategies. At the interface level, filtering, collection handling, and export were redesigned to reduce interaction friction. At the workflow level, the thesis investigated whether LLM-assisted pre-population could reduce the burden of manual submission while retaining user review and control. In this sense, Iteration 1 does not itself demonstrate that LLM assistance is the correct solution; rather, it demonstrates that manual submission is a sufficiently serious usability problem to justify evaluating automation as a design response.

5.5 Usability Evaluation Iteration 2

This section reports the second usability evaluation of InterDev, conducted after Iteration 1 identified friction in filtering, collection management, data export, and manual impact evaluation submission. Iteration 2 introduces three interface- and workflow-level extensions: multi-criteria filtering, multi-collection curation with improved export affordances and formats, and an LLM-assisted submission pathway that pre-populates metadata from uploaded evaluation reports while retaining user review prior to ingestion. A between-groups design compares manual submission against LLM-assisted submission using GPT-4o, Gemini 1.5 Pro, and Claude 3.5 Sonnet. Usability was assessed using task completion measures and the Post-Study System Usability Questionnaire (PSSUQ), alongside perceived accuracy and trust ratings for the submission workflow. Ethics for this study was approved by the Pennsylvania State University Internal Review Board (IRB), as the evaluation was undertaken whilst on a PhD internship placement.

5.5.1 Updated User Requirements

Building on the formative findings from the first evaluation, the following user requirements were refined based on specific issues identified in the user testing results:

Requirement 1: Enhanced Filtering and Contextualisation

The first usability evaluation highlighted the need for more intuitive filtering mechanisms. In Task 4, where users were asked to filter impact evaluations based on sector, significant variability in task completion times was observed, suggesting difficulties in locating and applying filters efficiently. Additionally, thematic analysis revealed that 13 user comments referenced difficulties in Guidance and Navigation, emphasizing the need for clearer filtering processes. In response, the updated InterDev platform refines the filtering experience by allowing dynamic, multi-criteria filtering, improving UI elements for filter selection to enhance usability, and implementing additional filtering categories based on user feedback.

Requirement 2: Improved Data Access and Download Mechanisms

Users in the first evaluation expressed confusion regarding the data download process, with Task 7 showing the highest number of intervention requests. Thematic analysis further indicated that Efficiency and Performance issues (9 mentions) were associated with this functionality. To address this, InterDev has introduced clearer UI indicators for data export and support for multiple download formats (e.g., CSV, JSON, and TTL).

Requirement 3: Multi-collection management

Feedback from the first usability study indicated a need for users to manage multiple collections of evaluations efficiently. In Task 5, which required adding evaluations to a collection, participants experienced higher median completion times, with multiple intervention requests recorded. Usability and Interface Design (UID) was the most frequently mentioned theme (19 instances), with specific references to difficulties in collection management. The updated version now includes the ability to create and manage multiple collections, improved UI elements for organising, sharing, and renaming collections, and enhanced metadata tagging for better collection categorization.

Requirement 4: Improved Data Submission Process

The first usability evaluation revealed significant challenges in the manual data submission process, particularly in Task 8 (submitting a new impact evaluation) and Task 9 (finding the submitted evaluation in the evidence view). Participants encountered difficulties in structuring data correctly, resulting in multiple intervention requests. The thematic analysis highlighted 12 mentions of Functionality and Features (FF) concerns, with users suggesting

that automation or guidance would improve submission efficiency. To address these issues the submission workflow was updated with clearer guidance and validation steps, most notable with the introduction of LLM-Assisted data submission and validation to support users in structuring evaluation data, and enabled user verification of LLM-generated RDF before finalising submission.

5.5.2 Evaluating LLM Assisted Knowledge Graph Generation from Text

The first usability evaluation revealed significant challenges in the manual data submission process, particularly in Task 8 (submitting a new impact evaluation) and Task 9 (finding the submitted evaluation in the evidence view). Participants encountered difficulties in structuring data correctly, resulting in multiple intervention requests. The thematic analysis highlighted 12 mentions of Functionality and Features (FF) concerns, with users suggesting that automation or guidance would improve submission efficiency. These tasks also had longer completion times. This motivated the exploration of interventions that could reduce submission effort and improve data quality. Several alternative approaches were considered:

Interface-led guidance. A first class of alternatives involves improving the submission interface without automating extraction: clearer field definitions, examples and tooltips, progressive disclosure, wizard-based steps, and stronger inline validation. These interventions can reduce navigation errors and improve learnability, but they do not remove the dominant cost observed in Iteration 1: the manual transcription and interpretation required to map report content into structured metadata.

Controlled inputs and structured templates. Another option is to shift structure upstream by requiring contributors to provide metadata via templates (e.g., CSV/JSON forms) or standardised reporting checklists. While this can improve consistency, it is poorly suited to retrospective ingestion of legacy evaluation reports and increases adoption friction by requiring additional work before a contribution can be made.

Rule-based and conventional NLP extraction. A third approach is to use heuristics, regular expressions, or conventional information extraction pipelines (e.g., named entity recognition and relation extraction) to populate parts of the submission form. Such approaches can be effective for stable fields (e.g., DOI detection) but typically require

substantial domain- and format-specific engineering to handle the variability of evaluation reports, and they tend to degrade as document styles change.

Supervised document extraction models. Trainable document information extraction models offer stronger generalisation within a target distribution, especially when layout and field structure are consistent. However, applying them to ontology-aligned extraction across heterogeneous reports would require extensive labelled data capturing entities, relations, and ontology mappings, which was not available at this stage of the research.

Human curator workflows. Finally, the system could reduce end-user burden by assigning submission to trained curators. While likely to improve consistency, this approach does not scale well and conflicts with the aim of enabling broader contribution to the evidence graph.

Given these alternatives, the author of this thesis prioritised an approach that (i) reduces transcription burden rather than only improving interface clarity, (ii) can operate over heterogeneous, unstructured evaluation reports, (iii) can be aligned with ERCT’s ontology, and (iv) supports rapid iteration within the evaluation programme. This motivated the exploration of LLMs as a mechanism for automatic pre-population combined with human review. The following section therefore evaluates an LLM-assisted pipeline for extracting ERCT-aligned RDF from evaluation reports, while explicitly assessing its limitations (e.g., syntactic validity) and its implications for later usability and trust-focused evaluations. Additional research is presented in this section to address this challenge by introducing a pipeline that automates the process using state-of-the-art LLMs, including GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro. These models parse PDFs of evaluation reports, generating structured knowledge representations that align with the ERCT ontology. The generated outputs are compared against manually curated ground truth data using BLEU and Macro-F1 scores (Papineni et al., 2002; Sokolova & Lapalme, 2009) to evaluate and benchmark their performance, ultimately enhancing InterDev’s ability to integrate and utilise new evaluation data.

5.5.2.1 Context: Automated KG Construction and Document Parsing

The Automated generation of knowledge graphs (KGs) from unstructured textual data has garnered significant attention in recent years (Melnik et al., 2022). Traditional KG

construction relies on manual curation or rule-based systems, which are time-consuming and often lack scalability (Dong et al., 2014). Recent advancements in natural language processing (NLP) and machine learning have enabled more efficient and automated approaches to KG generation, using RDF triples to ensure structured, machine-readable formats (Mihindukulasooriya et al., 2023).

Several notable frameworks tackle the challenges of extracting entities and relationships from unstructured documents to generate RDF data. Grapher introduces a multi-stage KG generation pipeline that separates node (entity) generation and edge (relationship) construction, achieving state-of-the-art performance on datasets like WebNLG and New York Times (Melnik et al., 2022). This approach uses pre-trained language models for entity extraction and avoids redundancies inherent in graph linearization approaches (Agarwal et al., 2020; Dognin et al., 2021).

Similarly, Doc-KG presents a methodology for constructing local knowledge graphs from unstructured documents and mapping them to external resources like Wikidata. The model integrates syntactic information to extract accurate triples, enhancing KG enrichment while maintaining alignment with existing knowledge bases (Salman et al., 2024).

Other efforts, such as AutoIE and Attend, Copy, Parse, focus on document-level information extraction rather than knowledge graph generation (Liu et al., 2024; Palm et al., 2019). AutoIE introduces a domain-specific framework for extracting entities and relationships from scientific documents. Its architecture integrates layout analysis, functional block recognition, and entity-relation extraction using models like SBERT, achieving high accuracy on specialised datasets but producing outputs that are not explicitly linked as knowledge graph triples. Similarly, the Attend, Copy, Parse (Palm et al., 2019) architecture provides an end-to-end solution for extracting structured information from document images. It bypasses word-level labelling requirements by attending to and parsing text directly from raw inputs, enabling extraction of fields such as dates and values from semi-structured documents such as invoices.

5.5.2.2 Approach

The experimental setup focuses on automating the generation of RDF-based subgraphs from evaluation reports, aligning them with the existing ERCT knowledge graph

introduced in Chapter 2. This began with using curated ERCT data, which has already been uplifted into RDF triples, representing entities and relationships such as interventions, methodologies, and measured outcomes (Matt Murtagh-White et al., 2023; Murtagh-White et al., 2024). These RDF triples act as the ground-truth data for evaluation purposes. To align these subgraphs with their respective evaluation documents, the original PDFs were identified on which the knowledge graph data is based using the DOI.

Using PDFMiner¹⁵, raw text content was extracted from each evaluation report, creating a dataset where each sample pairs the document's raw text with its corresponding RDF subgraph. This dataset provided the foundation for testing the automated pipeline, ensuring the input (PDF content) and expected output (RDF triples) are aligned. Using this method, a dataset of 700 evaluation text-subgraph pairs were generated for this evaluation.

To test the effectiveness of using LLMs for this task, the extracted raw text data was passed to three state-of-the-art models: GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro. Each model was prompted to generate RDF triples adhering to the ERCT ontology and formatting. To guide the LLMs in producing RDF outputs that are both accurate and structured, an example mapping and a predefined schema was provided as part of the input prompt. This setup allowed for a comparative evaluation of the models' performance in automating the transformation of unstructured text into knowledge graph subgraphs. The prompt for this approach can be seen in Figure 12.

¹⁵ <https://github.com/pdfminer/pdfminer.six>

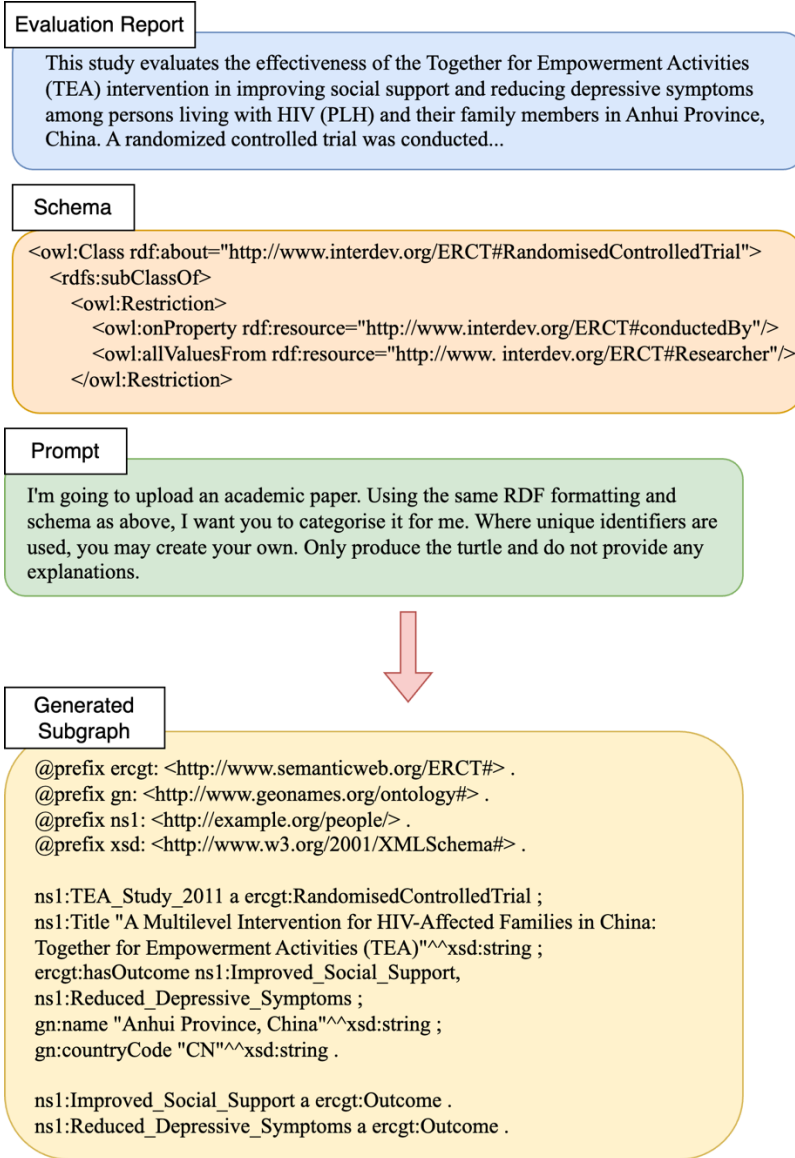


Figure 12: Prompt structure for extracting RDF triples from evaluation reports.

Once RDF triples were generated by each LLM, their quality was evaluated by comparing them to the manually curated ground-truth RDF subgraphs from the ERCT knowledge graph. The evaluation employed two widely used metrics: BLEU (Papineni et al., 2002) and Macro-F1 (Lin, 2004). BLEU scores assess the syntactic similarity between GPT-generated RDF triples and the ground truth by measuring n-gram overlap, while Macro-F1 scores evaluate precision and recall at the entity-relation level. Additionally, the generated subgraph was parsed using RDFLib¹⁶ to determine if they can be parsed as valid RDF,

¹⁶ <https://github.com/RDFLib/rdfliib>

testing compliance with RDF standards and the syntactical correctness of the generated outputs. The number of correctly parsed RDF subgraphs are reported as a percentage of the total.

5.5.2.3 Results

The results of the evaluation, which can be seen on Table 12, indicated that GPT-4o demonstrates strong performance in generating RDF triples from unstructured evaluation reports, achieving the highest BLEU score of 49.3 among the tested models. This suggests that GPT-4o can closely align its outputs with the expected structure and schema of the ERCT ontology. However, its Macro-F1 score of 30.9 is slightly lower than that of Gemini 1.5 Pro (32.4) and Claude 3.5 Sonnet (32.2), suggesting room for improvement in capturing semantic precision and recall for entity-relation extraction.

When evaluating the validity of the generated RDF triples using RDFLib, which checks if the Turtle snippets can be parsed as valid RDF, GPT-4o achieved a significantly lower validity rate of 11%, compared to 90% for Gemini and 26% for Claude. This discrepancy indicated that while GPT-4o excels in generating syntactically similar outputs, it struggles with adhering to RDF standards, leading to outputs that are less consistently parseable.

These results demonstrate a trade-off in GPT-4o's performance. Its higher BLEU score suggests strong lexical and structural fidelity to the schema, yet the low validity percentage indicates a need for enhanced model guidance or post-processing to ensure RDF compliance. In contrast, Gemini's combination of the highest Macro-F1 score and the highest validity rate demonstrates a more balanced approach, making it particularly effective for generating RDF triples that are both semantically accurate and structurally valid.

Model	BLEU	F1	%Valid
GPT4o	49.3	30.9	0.11
Gemini 1.5 pro	45.7	32.4	0.90
Claude 3.5 Sonnet	45.2	32.2	0.26

Table 12: Comparison of BLEU and Macro-F1 scores for RDF subgraph generation from evaluation reports

5.5.2.4 Conclusions

The research described in section 5.5.2 showed that LLMs can generate ERCT-aligned RDF subgraphs from unstructured evaluation reports, but with important differences in output quality across models. Empirically, GPT-4o achieved the highest BLEU score (49.3), indicating the strongest lexical and structural similarity to the target RDF. However, its outputs had a low RDF validity rate (11%), meaning that most generated snippets could not be parsed successfully. Gemini 1.5 Pro produced the highest Macro-F1 score (32.4) and the highest validity rate (90%), indicating a more balanced performance between semantic extraction quality and syntactic correctness. Claude 3.5 Sonnet performed between these extremes. Taken together, the results demonstrate that automated ERCT-aligned extraction is feasible, but that model choice materially affects whether outputs are merely similar to the target form or are actually usable within a knowledge-graph pipeline.

These findings have two direct design implications for InterDev. First, LLM-assisted submission is viable as a pre-population mechanism, but not as an autonomous ingestion pathway: generated outputs require validation before insertion and should remain subject to human review. Second, the results justify treating syntactic validity and reviewability as first-class design constraints in the submission workflow, rather than assuming that high textual similarity is sufficient for practical reuse. In other words, the main lesson is not simply that LLMs can help, but that they must be integrated through a controlled pipeline that balances LLM assistance with validation and user oversight. These design implications suggest that LLMs should not replace users in the submission pipeline, but that they can be used to pre-populate structured metadata while preserving user review prior to ingestion. In other words, the findings motivate a human-in-the-loop submission design

This section therefore motivates the next stage of evaluation in two ways. At the technical level, it establishes that LLM-assisted extraction can reduce the burden of manual structuring, but only when embedded in a guarded workflow. At the interface level, it motivates testing whether such pre-population actually improves user experience in practice, and whether the resulting gains in efficiency are achieved without undermining

user understanding or control. The subsequent usability evaluation addresses that question by comparing manual and LLM-assisted submission workflows under user study conditions.

5.5.3 Changes to System Design

The second usability evaluation of the InterDev platform was conducted following enhancements based on the insights gained from the first evaluation. These updates aimed to improve usability, streamline workflows, and integrate automated data submission functionalities powered by LLMs. Building on the stable architecture and interaction abstractions defined in Chapter 4, Iteration 2 introduced the following interface- and workflow-level extensions:

5.5.3.1 Filtering and Navigation Enhancements (Requirement 1)

To improve filtering and contextualisation, the Evidence View was updated with dynamic multi-criteria filtering. Users can now apply multiple filters simultaneously, such as country, sector, and methodology, and filter selection controls were redesigned to reduce friction and ambiguity as in Figure 13.

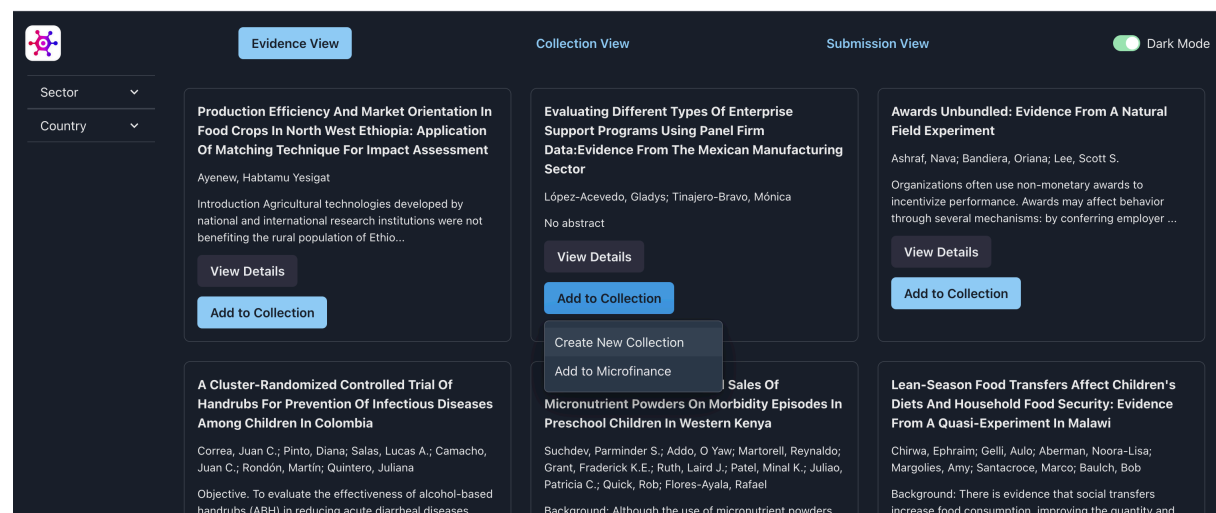


Figure 13: Iteration 2 Evidence View with multi-criteria filtering and collection actions.

5.5.3.2 Data Download Process (Requirement 2)

Enhancements were made to the data export functionality to address user difficulties with downloading evaluation data. The updated system provides clearer UI indicators for available export options, allowing users to download data in multiple formats, including

CSV, JSON, expanding on the previous iteration which only allowed for the export of TTL. The export data feature remains in the collection view.

5.5.3.3 Collection Management Features (Requirement 3)

To support more usable multi-collection management, a redesigned Collection View was introduced, as can be seen in Figure 14. Users could now create, manage, and download multiple collections, providing structured overviews of saved evaluations. The first iteration supported one unnamed collection only. Additional UI enhancements include improved organization, renaming capabilities, and metadata tagging to facilitate categorization and retrieval.

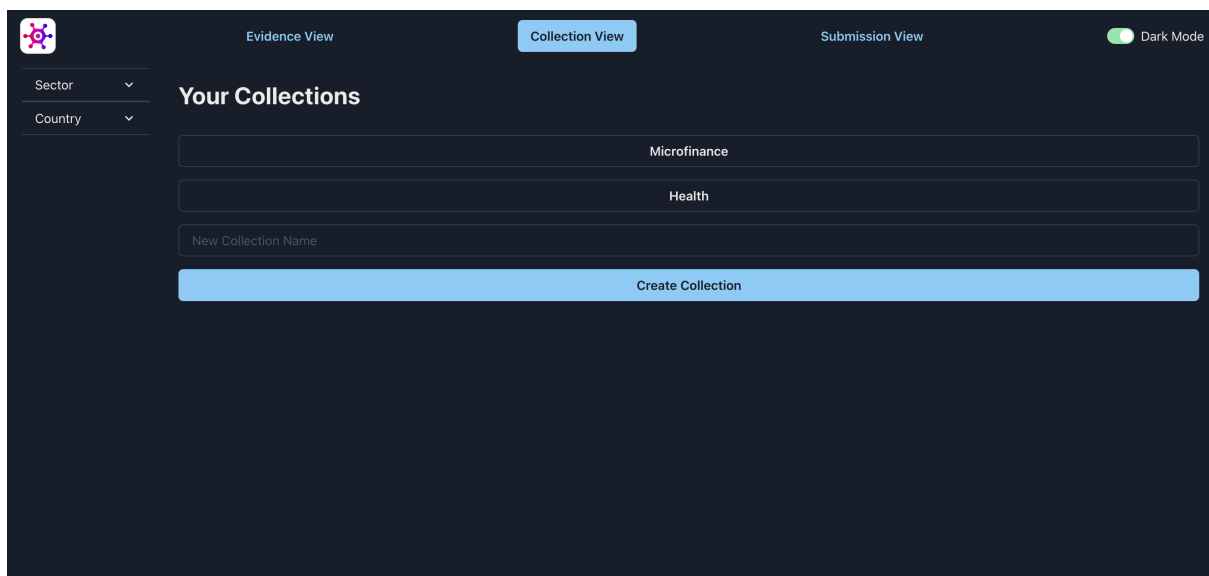


Figure 14: Iteration 2 Collection View, enabling multi-collection management.

5.5.3.4 Impact Evaluation Submission Workflow Updates (Requirement 4)

The impact evaluation submission process was significantly overhauled to address user challenges with manual data entry. A new impact evaluation Submission View was introduced, allowing users to upload evaluation reports, select an LLM model for automated RDF extraction, and review extracted pre-populated RDF triples as pre-filled form entries before finalizing submission as per Figure 15. New validation steps guide users in structuring evaluation data correctly in an effort to improve the efficiency and accuracy of data entry. The baseline version of the impact evaluation submission form exists from iteration one, where users opt to use manual submission without the use of pre-population via an LLM, albeit with the interface updates introduced in iteration two.

The screenshot shows the 'Trial Submission Form' in 'Submission View'. At the top, there are tabs for 'Evidence View', 'Collection View', and 'Submission View' (which is active). A 'Dark Mode' toggle is on the right. A green checkmark message states 'PDF processed and parsed successfully!'. Below this, the 'Upload PDF' section shows a file named '10.1186_s12889-016-3129-3.pdf'. The 'Select AI' section has a dropdown for 'AI Provider' set to 'GPT'. The 'Basic Info' section contains two text areas: 'Title' with the text 'The way to a man's heart is through his stomach?: A mixed methods study on causal mechanisms through which cash and in-kind food transfers decreased intimate p' and 'Abstract' with a longer paragraph about intimate partner violence (IPV) in Ecuador.

Figure 15: Iteration 2 Submission View with LLM-assisted pre-population from an uploaded evaluation report.

5.5.3.5 LLM-Assisted Data Submission Pipeline

A key extension introduced in iteration 2 is an LLM-assisted submission workflow that replaces the baseline manual metadata entry process described in Chapter 4 and evaluated in iteration 1 with a pre-populated, reviewable submission pipeline. The aim is to reduce manual entry effort while preserving a human-in-the-loop confirmation step prior to ingestion. The code for iteration 2 is available on the thesis GitHub, under the iteration 2 folder¹⁷.

Text Extraction from PDFs

Users upload an evaluation report in PDF format. The backend extracts raw text using PDFMiner¹⁸ (version 20250506) and passes the extracted content to the submission pipeline.

¹⁷ <https://github.com/mmurtaghw/InterDevResearch>

¹⁸ <https://pdfminersix.readthedocs.io/en/latest/>

LLM-assisted submission via field extraction.

Participants are assigned to one of four submission conditions: GPT-4o, Gemini 1.5 Pro, Claude 3.5 Sonnet, or a manual baseline with no model assistance. In the LLM conditions, the assigned model processes the extracted text and produces structured metadata required by the Submission View (e.g., intervention details, study design/methodology, outcomes, and setting/context). This step replaces manual transcription of report content into submission fields.

Ontology-aligned serialization and syntactic validation

The extracted metadata is serialized into ERCT-aligned RDF for storage using the same data model described in Chapter 4. The system then validates the serialized output prior to submission. In Iteration 2, this validation step is applied specifically to LLM-generated outputs to detect syntactic errors (e.g., invalid Turtle) before they reach the datastore. If validation fails, the submission is blocked and the user is shown an error indicating that the generated output could not be parsed.

User review and correction via form-based editing

Because users are not expected to edit RDF directly, the system presents the extracted metadata as editable form fields rather than raw triples. Users can correct or complete fields before final submission. After edits, the system re-serializes the revised fields and re-runs validation to ensure the final submission remains syntactically valid.

Submission and ingestion.

Once the user confirms the edited metadata and validation succeeds, the submission is ingested through the existing insertion pathway.

Overall, Iteration 2 introduces an LLM-assisted, form-based pre-population workflow for submission, coupled with pre-ingestion validation and user review, to reduce manual entry effort while maintaining user confirmation of the final metadata. This workflow enables a controlled comparison between manual submission and model-assisted pre-population, allowing Iteration 2 to measure effects on task success, efficiency, and user perceptions of accuracy and trust.

5.5.4 Ontology changes as a result of iteration 1

Iteration 1 exposed that several interaction breakdowns were not solely interface problems but reflected mismatches between user mental models and the ontology’s operational role in the system. Filtering and collection work was slowed by ambiguous facet naming and inconsistent semantics, while submission errors were repeatedly linked to uncertainty about what particular fields meant and when they should be used. These findings motivated treating the ontology as part of the interaction contract, meaning it needed to provide stable, human-readable terminology and a predictable facet structure that the UI could rely on during browsing, filtering, and data entry.

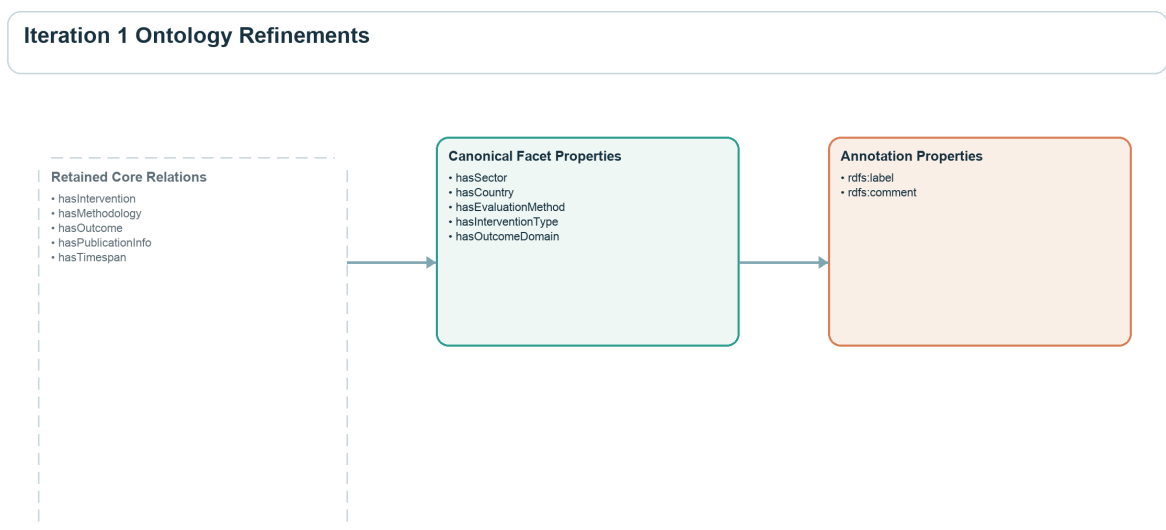


Figure 16: Iteration 1 ERCT ontology refinements, showing the retained core relations alongside the new canonical facet properties and annotation properties used to stabilise browsing, filtering, and interface terminology.

In response, the ERCT ontology was revised to better support interaction and reduce ambiguity at points of use. Naming and documentation were strengthened so that the UI could surface consistent, user-facing terminology. Missing `rdfs:label` values were added across core classes and properties, and short clarifying definitions were provided using `rdfs:comment` to make intent and expected usage explicit. V3 also introduced lightweight UI-hint annotations, namely `erct:uiSection`, `erct:uiPlaceholder`, and `erct:uiExample`, enabling the interface to provide structured guidance in high-friction fields without requiring custom, hard-coded UI logic. To address filtering breakdowns directly, a canonical facet model was introduced for UI configuration, mapping

common filters to stable properties over controlled concept IRIs, including `erct:hasSector`, `erct:hasCountry`, `erct:hasEvaluationMethod`, `erct:hasInterventionType`, and `erct:hasOutcomeDomain`. These facet properties are now modelled to use controlled IRIs to support consistent filtering and reuse; however, legacy data may still encode some facet values as literals, meaning alignment or migration may still be required within ingestion and curation pipelines to fully normalise older records. Ambiguous or misspelled legacy terms that undermined consistency were deprecated in place using `owl:deprecated true` with migration notes, preserving backward compatibility while steering new data toward canonical properties.

5.5.5 Experimental Design

To assess the usability and effectiveness of these new features, this study follows a hybrid qualitative-quantitative approach. The evaluation methodology consists of task-based usability testing, between-groups testing for LLM-assisted submission workflows, and user surveys. The study consists of structured task completion tests and system usability questionnaires (PSSUQ) to gather data. The ethics approval for this study was granted by the Institutional Review Board (IRB) at The Pennsylvania State University under the protocol STUDY00026713¹⁹. Participants were given a \$20 gift voucher for participation in this study.

¹⁹ This work was undertaken during a 6 month research placement at The Pennsylvania State University.

5.5.5.1 Experimental Design Modifications

Iteration 2 adopted a more evaluative study design than Iteration 1. Whereas Iteration 1 used supervised think-aloud to diagnose interaction breakdowns and generate redesign targets, Iteration 2 focused on comparative validation using task outcomes and structured post-task measures. Accordingly, Iteration 2 employed a between-subjects design to evaluate whether LLM-assisted submission improves performance and user-perceived quality relative to manual submission, while reducing the risk of learning effects that would confound within-subject comparisons (Shadish et al., 2002).

Participants were assigned to one of four conditions: a manual submission control condition (no LLM assistance) or one of three LLM-assisted submission conditions (GPT-4o, Gemini 1.5 Pro, or Claude 3.5 Sonnet). This structure supports a comparison of assisted versus manual submission while controlling for a degree of model-specific behaviour.

5.5.5.2 Participant Setup and Recruitment

Iteration 2 was conducted as an unsupervised remote study. Participants completed the procedure independently on a desktop or laptop via a study link and were guided through the workflow using an online survey instrument (Qualtrics), which provided task prompts, condition assignment, and post-task questionnaires. Because the study was unsupervised, no screen or audio recordings were collected and no real-time assistance was provided during task execution.

To support standardised outcome measurement under these conditions, task completion time and completion status were captured using survey timestamps and system logs. Subjective measures were collected immediately after task completion, including PSSUQ, a perceived accuracy rating for the submission output, and an overall trust rating.

Each participant completed six tasks: five core browsing and filtering tasks (selecting evaluations, applying single and combined filters, building collections, and downloading collections), followed by a metadata submission task performed under their assigned condition. Participants in the manual condition entered metadata fields directly, while participants in the LLM-assisted submission conditions reviewed and corrected pre-

populated fields generated by their assigned model. The procedure concluded with the 19-item PSSUQ.

Participants were recruited using the same target criteria as Iteration 1, focusing on researchers with a social science background and/or experience or interest in international development evidence. Recruitment was conducted via email outreach and institutional advertising at Pennsylvania State University (e.g., posters within the School of Information Science and Technology). In total, 20 participants completed the study.

5.5.5.3 Task List and Evaluation Instruments

Table 13 summarises the task sequence used in Iteration 2 and shows how each task operationalises the updated user requirements (R1–R4). Tasks 1–2 exercise evidence discovery and multi-criteria filtering (R1). Tasks 3–5 focus on multi-collection curation and export workflows (R2–R3), including verification of collection contents and download in alternative formats. Task 6 evaluates the revised submission workflow under manual and LLM-assisted submission conditions (R4), capturing both interaction outcomes (completion, time) and post-task perceptions of accuracy and trust.

Step	Task	Requirement
1	Use the Evidence View to locate an evaluation conducted in a given country and note its sector.	R1
2	Return to the Evidence View and apply a sector filter to display only impact evaluations from that sector.	R2
3	Create a new collection and add at least four impact evaluations from the filtered list.	R3
4	Apply both a country and a sector filter, then add up to four more impact evaluations to the collection.	R1, R3
5	Confirm the contents of the collection in the Collection View and download it in the preferred format.	R2, R3
6	In the Submission View, upload an evaluation report (or enter an identifier), complete the metadata entry (LLM-assisted or manual), and submit the impact evaluation.	R4

Table 13: Updated task list and associated requirements for second evaluation.

5.5.6 Results

This section reports the quantitative outcomes of Iteration 2, organised by instrument and workflow stage. Section 5.5.6.1 summarises perceived usability using PSSUQ, Section 5.5.6.2 reports task completion times for the browsing and curation tasks, and Sections 5.5.6.3–5.5.6.4 report submission outcomes and perceptions (completion, duration, perceived accuracy, and trust) under manual versus LLM-assisted submission conditions. The anonymised raw study data are available in the InterDev repository under the iteration 2 folder²⁰.

5.5.6.1 PSSUQ

The Post-Study System Usability Questionnaire (PSSUQ) was used to evaluate participants' overall satisfaction with the InterDev platform following task completion. Figure 17 presents the distribution of ratings across all 19 questionnaire items and the three standard subscales: System Usefulness, Information Quality, and Interface Quality. Lower scores indicate better usability.

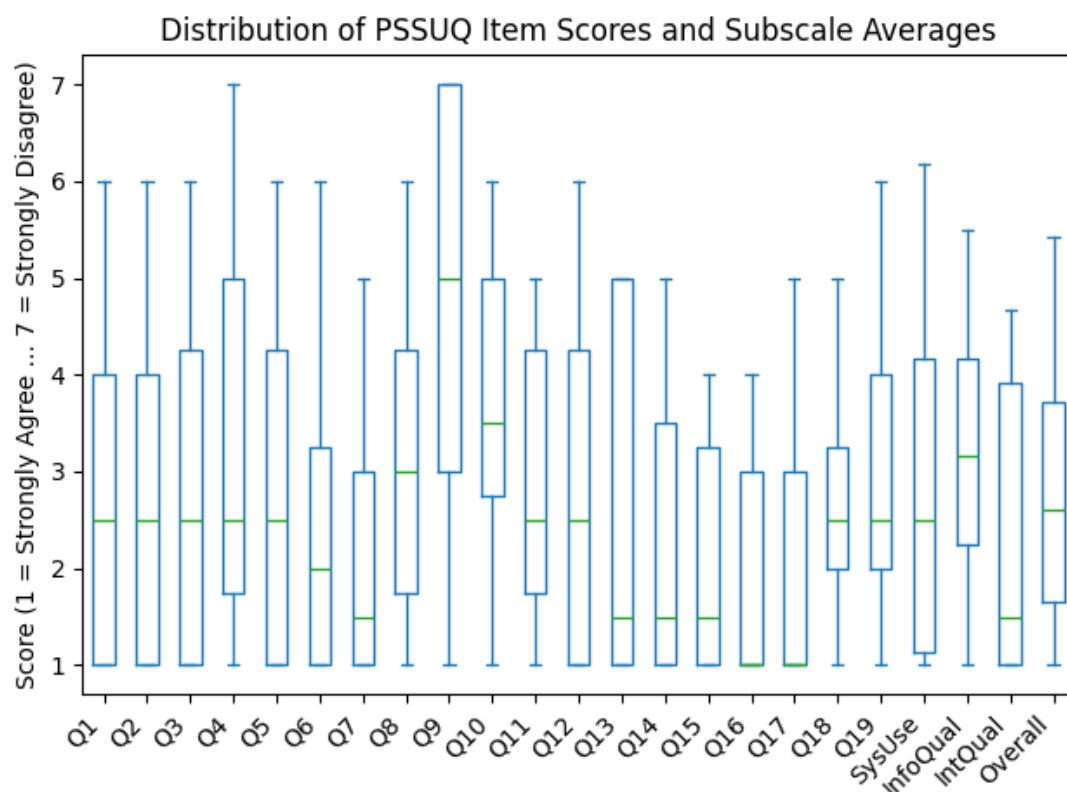


Figure 17: Distribution of PSSUQ Item Scores and Subscale Averages

²⁰ <https://github.com/mmurtaghw/InterDevResearch>

The overall median score was approximately 2.6, which lies within the “Agree” range, reflecting a generally positive user experience. The System Usefulness subscale (items Q1–Q6) achieved a median of 2.4 with an interquartile range (IQR) from 2.0 to 3.0, indicating that most participants felt the platform effectively supported task completion. The Information Quality subscale (Q7–Q12) displayed greater variability, with a median of 2.8 (IQR 2.3–3.5). Item Q9, which concerns the clarity of error messages, produced several high outliers (scores of 5–6), suggesting that some users encountered unclear or unhelpful error feedback. Interface Quality (Q13–Q15) showed the strongest performance, with medians between 1.8 and 2.1 and narrow IQRs, reflecting agreement that the interface was visually clear and easy to learn.

At the individual item level, Q5 (“I was able to efficiently complete the tasks and scenarios”) had a relatively high median of 3.2 and a wide spread, indicating inconsistent perceptions of efficiency across participants. Learnability (Q7) exhibited a low outlier at 1, showing that a small number of users experienced difficulty when first interacting with the system. Wider distributions for Q11 and Q12 suggest that some participants had difficulty locating specific information. These findings indicate that while overall usability is strong, further enhancements in error communication, task efficiency, and early learnability remain necessary.

5.5.6.2 Task Timing

Figure 18 displays task completion times for the five browsing and curation tasks. Task 1 (selecting an impact evaluation from the Evidence View) had a median completion time of approximately 73 seconds with an IQR of 50–85 seconds and several high-end outliers up to 145 seconds. This variability reflects exploratory behaviour among first-time users. Task 2 (applying a sector filter) showed far greater consistency, with a median of 27 seconds and minimal dispersion (IQR 18–38 seconds), indicating rapid user adaptation to filtering mechanics.

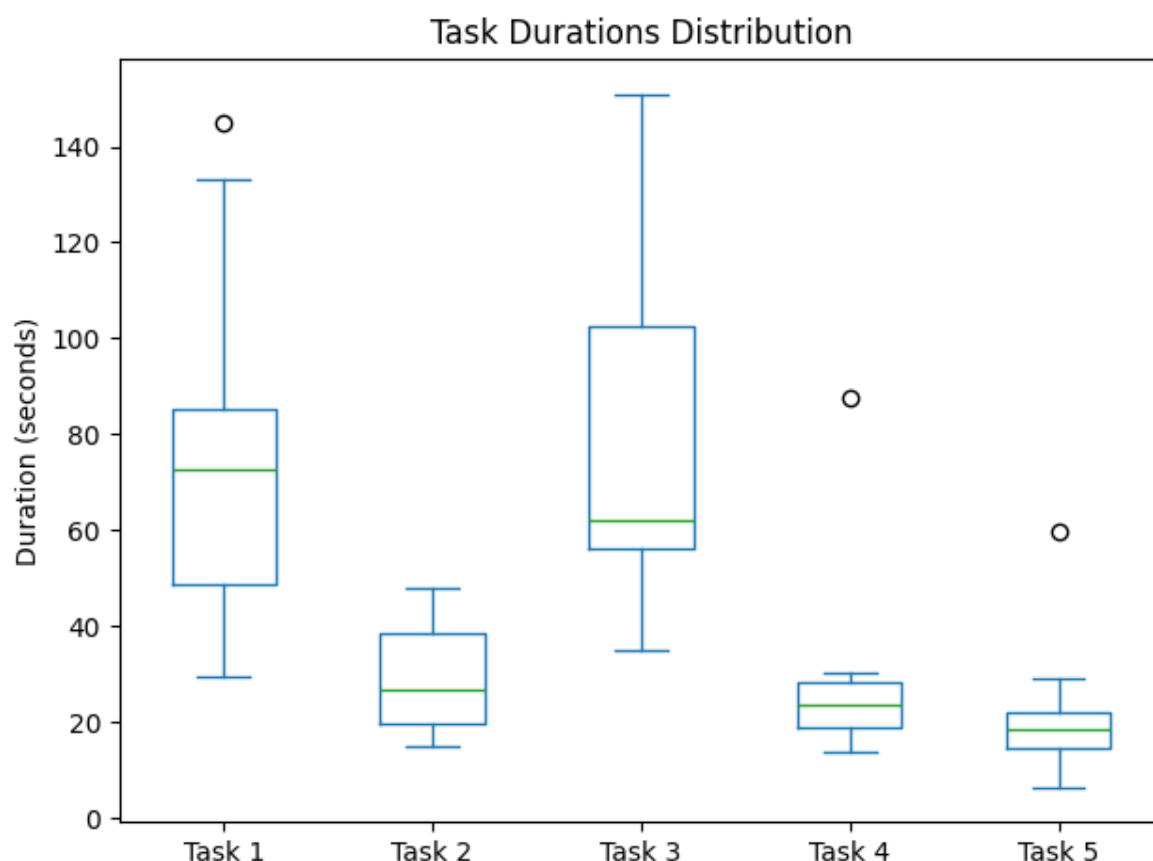


Figure 18: Task Durations Distribution

Task 3 (creating a collection and adding impact evaluations) required more time, with a median of 62 seconds (IQR 55–100 seconds), reflecting its greater complexity. Tasks 4 and 5 were completed most quickly and consistently. Task 4 (applying combined filters) had a median of under 25 seconds (IQR ~6–8 seconds), and Task 5 (downloading a collection) showed a similar completion time. These results demonstrate that once users understood the filtering and interaction model, they were able to complete subsequent tasks efficiently.

5.5.6.3 LLM Assisted Submission

Table 14 compares completion rates for the metadata submission task (Task 6) between the LLM-assisted submission and manual submission workflows. The LLM-assisted submission condition achieved a completion rate of 100 percent, while only 56 percent of participants in the manual condition completed the task. This suggests that the manual workflow introduced barriers that led to user frustration or task abandonment, whereas LLM assistance helped maintain progress and task success.

Metric	LLM-Assisted	Manual
Mean Duration (s)	227.45	496.90
Completion Rate	1.00	0.56
Mean Accuracy Rating	3.90	2.89
Trust in Submission System	3.20	2.33

Table 14: Metrics for LLM-Assisted Submission and Manual Workflows Submission

LLM-assisted submission also had an impact on efficiency. As shown in Table 14, the average task duration for metadata submission in the LLM-assisted submission condition was approximately 227 seconds, compared to roughly 496 seconds for the manual condition. This represents a reduction of more than 50 percent in task time. The increased efficiency is likely due to automated metadata extraction reducing manual entry effort and lowering cognitive load. No notable differences were observed between the three LLMs considered in the evaluation.

5.5.6.4 User Trust and Perceived Accuracy

Participants were asked to rate the perceived accuracy of the metadata produced in Task 6 on a seven-point scale. As shown in Table 14, the LLM-assisted submission workflow received a mean perceived accuracy score of 3.9 (between “Slightly accurate” and “Moderately accurate”), while the manual workflow averaged 2.9. These results suggest that users had higher confidence in metadata generated with LLM support compared to entirely manual entry.

Table 14 reports trust ratings for the submission workflow. Participants in the LLM-assisted submission group reported higher trust overall, with responses centred between “Moderately” and “Mostly,” and no scores below “Slightly.” The manual group centred on “Slightly,” with several participants reporting “Not at all.” The narrower IQR for the LLM group indicates more consistent trust experiences. These findings suggest that LLM integration not only improves efficiency and completion rates but also enhances user confidence in the workflow.

5.5.7 Limitations

Iteration 2 provides stronger evaluative evidence than the iteration 1, but several limitations constrain the interpretation of the results. The sample size was larger but still modest ($n = 20$), limiting statistical power.

The study was conducted remotely and unsupervised, which improves scalability but reduces experimental control. Task durations and completion status were derived from logs and survey timestamps, which may include off-task time or be affected by interruptions, technical issues, or variations in participant engagement. Without screen recordings or concurrent verbalisation, it is not possible to localise specific breakdowns or to distinguish confidently between usability problems and external distractions.

Finally, the evaluation focused on efficiency, completion, and subjective perceptions of accuracy and trust, but it did not directly assess the semantic correctness or downstream utility of the submitted metadata. Perceived accuracy and trust are imperfect proxies for data quality and may be influenced by automation bias or by the fluency of pre-populated forms. Consequently, these results should be read as evidence about interaction cost and user acceptance, not as a guarantee of improved knowledge graph quality.

5.5.8 Iteration 2: Conclusions

Iteration 2 provides empirical evidence that the revised submission workflow reduced interaction cost relative to the manual baseline within this study. Participants in the LLM-assisted submission conditions completed the submission task more often, completed it more quickly, and reported higher perceived accuracy and trust than participants in the manual condition. These findings support the narrower claim that, under the conditions tested here, LLM-assisted pre-population can reduce practical barriers in metadata submission and make contribution workflows easier to complete.

At the same time, the study does not show that LLM-assisted submission improved the semantic correctness of submitted metadata or the downstream quality of the knowledge graph. The main outcomes in this iteration are efficiency- and perception-oriented: completion, time, perceived accuracy, and trust. These are useful indicators of interaction cost and user acceptance, but they are not direct evidence that participants verified the

generated metadata carefully or that the submitted content was more correct than in the manual condition.

From a design perspective, the results suggest two implications. First, pre-populated, reviewable submission workflows are a promising way to lower the burden of contributing structured evidence, particularly where manual entry is costly and abandonment-prone. Second, because improved completion and higher perceived trust may also increase the risk of uncritical acceptance, LLM assistance should remain bounded by validation, user review, and inspectable provenance rather than being treated as autonomous submission. In other words, Iteration 2 supports using LLMs as a mixed-initiative contribution aid, not as a replacement for human judgement.

Cross-iteration interpretation should remain cautious. Iteration 2 was run remotely and unsupervised with a new participant cohort, whereas Iteration 1 was supervised and formative. Accordingly, the strongest claims from Iteration 2 are within-study claims about the relative effects of manual versus LLM-assisted submission in this evaluation. The broader implication is that once contribution friction is reduced, the evaluation focus must shift from efficiency alone to the epistemic consequences of LLM-mediated exploration: whether users verify outputs appropriately, how confidence is calibrated, and how transparency and provenance cues shape reliance during evidence use. This motivates the design and evaluation focus of Iteration 3.

5.6 Usability Evaluation Iteration 3

Recent work suggests that LLM-mediated exploratory search can change how users formulate questions and interpret retrieved material, but it can also introduce reliability risks, such as hallucinations, and higher cognitive demands in conversational settings (Reichert et al., 2022; White, 2024). These effects place pressure on users to verify, compare, and reconcile outputs rather than accept single responses at face value, raising concerns about confidence calibration and over-reliance (Gero et al., 2024). This reflects a recurring interaction trade-off: reducing friction through more fluent, engaging interfaces can inadvertently weaken reflective practices that support trust and careful judgment (Gero et al., 2024). Iteration 3 therefore evaluates InterDev under controlled interface conditions that vary only the mediation layer and the visibility of provenance cues, allowing analysis of workload, trust calibration, verification behaviour, and synthesis quality.

Iteration 2 demonstrated InterDev as a usable platform for evidence curation and contribution after addressing the primary issues observed in the first evaluation. Filtering and navigation became faster and more consistent once users understood the faceted interaction model, collection management better reflected realistic evidence workflows, and the export pathway was clarified through more visible download affordances and multiple formats. Most notably, introducing LLM-assisted submission substantially reduced task duration, increased completion rates, and improved perceived accuracy and trust relative to manual entry. Together, these results suggest that increased LLM-assisted submission can remove practical barriers to submission and reduce the interactional cost of structuring evaluation data.

However, Iteration 2 also exposed an important limitation in what usability and efficiency metrics can establish. Improvements in completion rates and satisfaction do not guarantee that a user understands, appropriately verify, or correctly interpret automatically generated outputs, particularly when the underlying data representation (e.g., RDF) is unfamiliar. The strong performance of the LLM-assisted submission workflow raises a plausible risk that reduced friction shifts user effort from inspection and sensemaking toward acceptance, especially when error feedback is unclear or when participants lack the domain-specific literacy needed to assess structured representations. In this sense, the core question shifts to a different paradigm of LLM-mediated interaction: from whether users can complete the LLM-assisted submission workflows to how LLM-mediated exploration reshapes cognitive effort, confidence, and verification behaviour during evidence use.

Iteration 3 therefore extended user evaluation beyond interaction success to examine LLM-mediated exploration during synthesis-oriented tasks. Rather than treating LLM support as a single “better or worse” feature, the system implements controlled interface conditions that vary only the mediation layer: baseline browsing without chat assistance, LLM-mediated exploration without explicit provenance cues, and LLM-mediated exploration with surfaced knowledge-graph-derived citations. Holding the evidence and query semantics constant allows observed differences to be attributed to interface-level transparency and mediation effects. This user evaluation design directly targets the concerns raised by Iteration 2: whether LLM-mediated exploration, whether users calibrate

confidence appropriately, and whether provenance cues support or alter verification practices.

5.6.1 Identified Requirements

Requirement 5: Transparency and Provenance for LLM-Mediated Exploration

Iteration 2 demonstrated that LLM-assisted submission substantially improves task completion and increases perceived trust in the system. However, this assistance also introduced an interpretability gap: users are not expected to possess working knowledge of RDF, yet they are required to review and approve LLM-generated structured representations before insertion into GraphDB. As a result, verification is displaced from a familiar textual domain into an abstract structural one, increasing the risk that users accept outputs they cannot meaningfully evaluate.

To address this limitation, Iteration 3 shifts verification away from low-level data structures and toward evidence provenance at the interface level. Rather than asking users to reason about RDF correctness, the system surfaces explicit links between system outputs and the underlying impact evaluations from which they are derived. By making citation visibility a controlled interface manipulation, Iteration 3 enables users to assess reliability through familiar scholarly practices: checking sources, comparing studies, and tracing claims to evidence, while also allowing systematic evaluation of how provenance cues influence reliance, confidence, and verification behaviour.

Requirement 6: Trust Calibration Rather Than Trust Maximisation

Iteration 2 showed higher mean trust and perceived accuracy in the LLM-assisted submission workflow. While this increase is encouraging, it is beneficial only insofar as trust is appropriately calibrated to actual correctness. Over-trust risks the silent propagation of structured errors into the knowledge graph, whereas under-trust risks abandonment of genuinely useful assistance. Trust, therefore, cannot be treated as a unidimensional success metric.

To address this, Iteration 3 explicitly distinguishes between subjective trust, self-reported confidence, and calibrated judgment. The evaluation design incorporates task-level confidence elicitation and formal calibration analysis, allowing confidence to be assessed

against objective performance rather than interpreted in isolation. By analysing confidence–correctness alignment under different interface conditions, Iteration 3 examines whether LLM-mediated interaction induces appropriate caution or encourages misplaced certainty, providing a more principled account of trust in LLM-mediated evidence systems.

Requirement 7: Reduced Cognitive Workload Without Sacrificing Judgment Quality

Iteration 2 primarily addressed efficiency and usability concerns, with LLM-assisted submission substantially reducing task duration and eliminating many completion failures. However, reductions in interaction time do not necessarily reflect reductions in cognitive effort. Efficiency gains may simply reallocate effort from data entry to sensemaking and verification, or in some cases suppress scrutiny altogether. As a result, time-on-task alone is insufficient to determine whether LLM-mediated interaction meaningfully supports evidence reasoning.

Iteration 3 introduces a requirement that LLM-mediated exploration should reduce unnecessary interaction burden without diminishing users’ capacity for careful judgement. This requirement is evaluated by measuring subjective workload directly, using NASA–TLX (Carroll & Rosson, 1987), alongside task performance and confidence measures. In this way, the study can distinguish between interface conditions that genuinely support reasoning and those that merely accelerate interaction while weakening reflective evaluation.

Requirement 8: Synthesis Support Over Curated Evidence

By the end of Iteration 2, the main interaction barriers to browsing, filtering, collection building, and export had been substantially reduced. As these curation mechanisms become more usable, the main challenge shifts from accessing individual evaluations to synthesising evidence across multiple studies. In practical terms, users must be able to compare findings, identify patterns or disagreements, and formulate coherent, evidence-based interpretations.

Iteration 3 therefore introduces a requirement that the system should support higher-order synthesis over curated evidence rather than only retrieval or isolated summarisation. To meet this requirement, LLM-mediated interaction is constrained to participant-defined

collections, ensuring that any generated synthesis is grounded in a set of evaluations intentionally selected by the user. This allows the evaluation to focus on how interface mediation affects comparison, synthesis, and interpretive confidence while preserving user control over the underlying evidence base.

5.6.2 Changes to System Design

Iteration 3 retains the system architecture, ERCT knowledge graph, and retrieval pathway described in Chapter 4. The changes introduced in this iteration are confined to the interaction layer and are designed to support controlled study of how LLM mediation and provenance visibility shape evidence exploration, comparison work, confidence, and verification behaviour.

5.6.2.1 Evidence View extensions introduced in Iteration 3

The Evidence View was extended to support richer discovery through (i) additional faceted browsing categories and (ii) keyword search over impact evaluation titles and descriptors.

These additions increase flexibility in locating relevant impact evaluations while preserving the same underlying retrieval semantics. Figure 19 illustrates the updated Evidence View, including the expanded facet panel (left) and the keyword search bar (top).

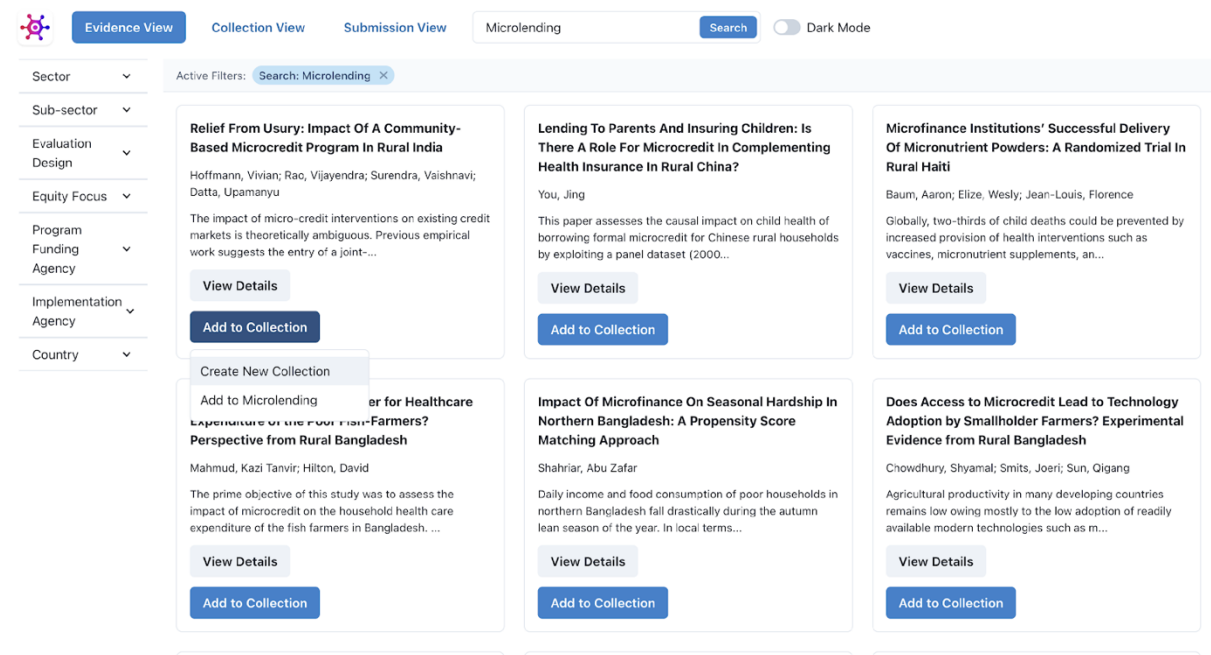


Figure 19: The evidence view of the third iteration of InterDev. The evidence view now contains additional categories for faceted browsing (left) and text search (top).

5.6.2.2 Multi-collection workflow as the unit of work.

Iteration 3 builds on the multi-collection capability introduced previously by positioning collections as the primary organisational unit for downstream comparison and synthesis tasks.

The Collection View provides access to multiple user-defined collections and supports navigation into a selected collection context for deeper interaction. Figure 20 shows the Collection View used in Iteration 3, where users can create and select collections that act as bounded working sets.

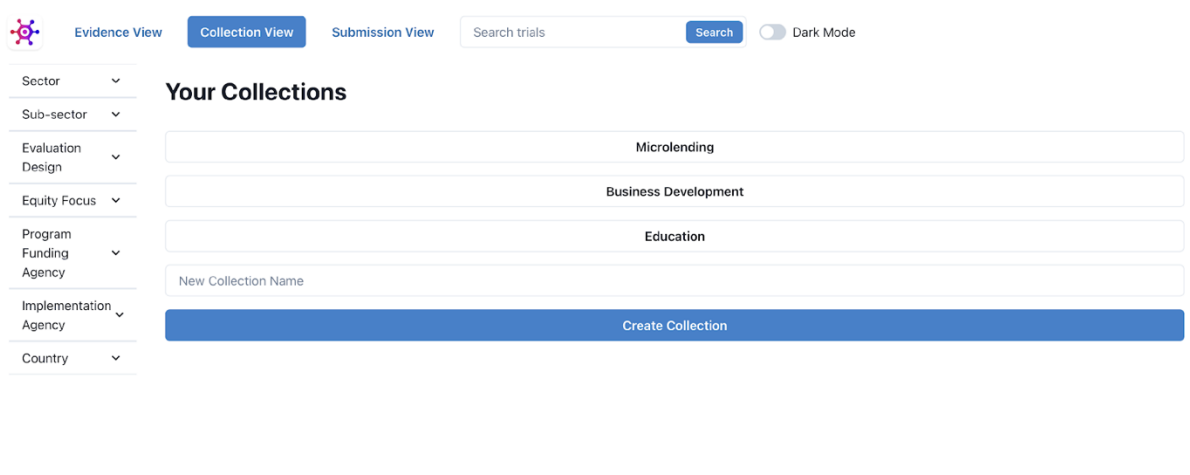


Figure 20: Collection View of the Third Iteration of InterDev. Collections are user defined subgraphs of evidence from the ERCT knowledge graph.

5.6.2.3 Collection-bounded LLM mediation and experimental condition

Iteration 3 introduces an optional LLM-mediated exploration layer that is available only within a collection context. This design enforces that any model responses are grounded in the evidence contained in the active collection rather than in the broader evidence base.

Figure 21 shows the in-collection interface for the chat-only condition, where the participant interacts with an LLM that answers questions by operating over the user-defined collection evidence.

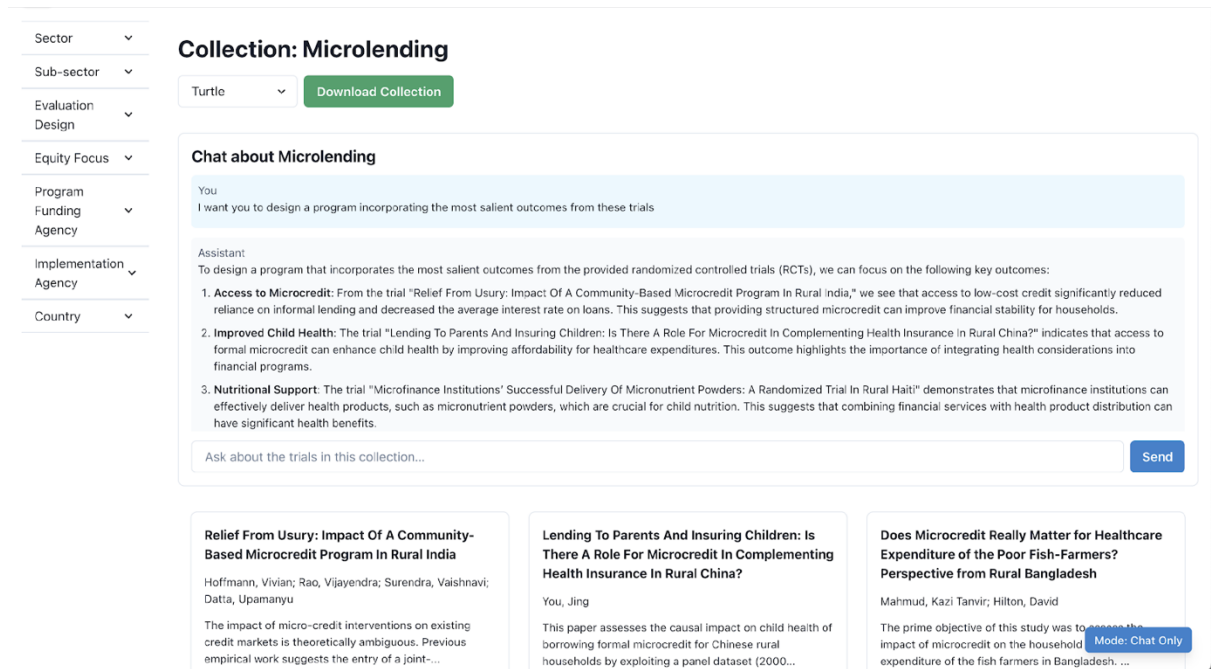


Figure 21: Inside a collection on the third iteration of InterDev. The "LLM-mediated exploration (no provenance cues)" condition includes an LLM that answers to user questions by retrieving data from the underlying user-defined subgraph.

5.6.2.4 Provenance-cue visibility manipulation

Three interface conditions are implemented: (C0) baseline browsing without chat assistance, (C1) chat based LLM-mediated exploration without explicit provenance cues, and (C2) chat-based LLM-mediated exploration with surfaced citations linking responses to specific impact evaluations, i.e. provenance cues. The conditions differ only in mediation and the visibility of provenance information, enabling attribution of observed differences to interface-level transparency rather than to changes in the underlying evidence available to participants.

5.6.3 Ontology Changes as a result of iteration 2

Iteration 2 introduced an LLM-assisted submission workflow and new export formats, which shifted the ontology requirements from supporting only browsing and manual entry to supporting auditable, mixed-initiative ingestion. Usability improvements in Iteration 2 showed that pre-population can reduce effort and increase completion rates, but the remaining outliers in error clarity and inconsistent perceptions of efficiency pointed to a need for stronger, model-level support for validation, provenance, and predictable field

types. In response, the ERCT ontology was updated to make submission and extraction events first-class, to improve the interface’s ability to render correct controls and messages for high-friction fields, and to reduce ambiguity in filtering and reuse by tightening how facets are represented.

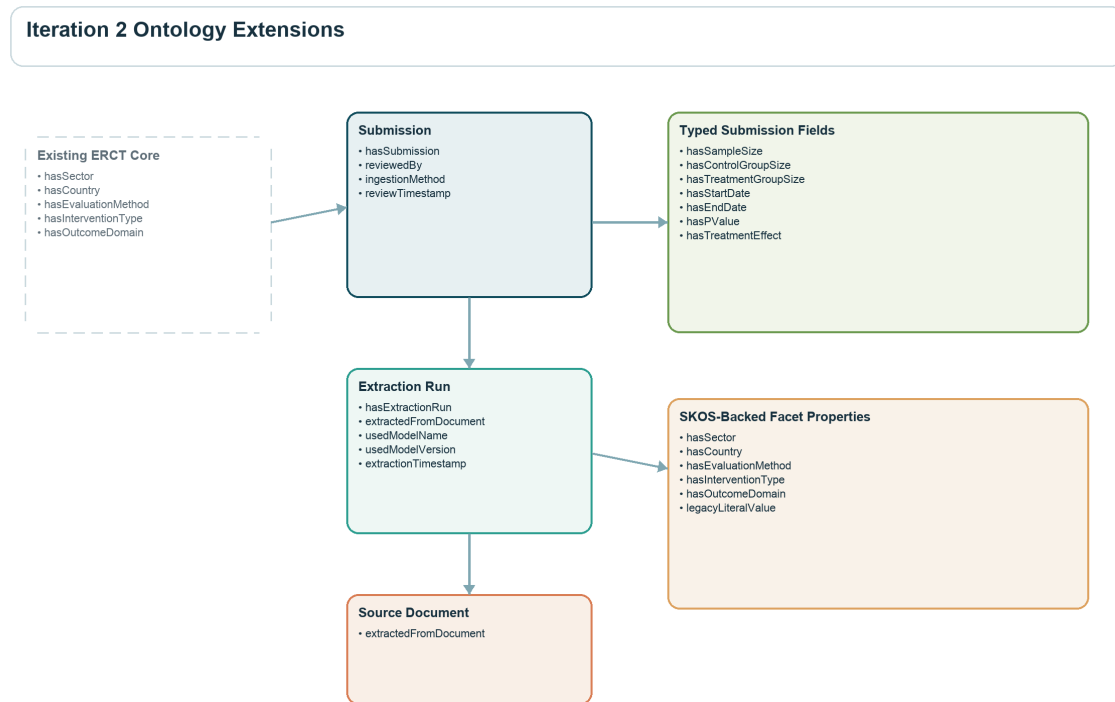


Figure 22: Iteration 2 ERCT ontology extensions, showing the submission and extraction provenance properties, typed submission fields, source-document linkage, and SKOS-backed facet properties added to support LLM-assisted ingestion and controlled reuse.

The updated version extends ERCT in three main ways while preserving backward compatibility. First, it introduces explicit provenance for ingestion and LLM-assisted submission by adding `erct:Submission`, `erct:ExtractionRun`, `erct:SourceDocument`, and `erct:IngestionMethod`, along with `erct:hasSubmission`, `erct:hasExtractionRun`, `erct:extractedFromDocument`, `erct:reviewedBy`, and `erct:ingestionMethod`, plus supporting timestamps and model descriptors such as `erct:extractionTimestamp`, `erct:reviewTimestamp`, `erct:usedModelName`, and `erct:usedModelVersion`. This aligns submission outputs with a traceable provenance record that can be surfaced in the UI and exported for audit. Second, it strengthens typing for submission fields by making domains and ranges explicit for key numeric and temporal properties, including `erct:hasSampleSize`, `erct:hasControlGroupSize`, and `erct:hasTreatmentGroupSize` as

`xsd:integer`, and reinforcing decimal and date typing for measures and timelines used during entry and export. Third, it formalises faceted exploration around controlled concepts by importing SKOS and shifting canonical facet properties to `skos:Concept` ranges, with scheme scaffolding via `erct:SectorScheme`, `erct:CountryScheme`, `erct:EvaluationMethodScheme`, `erct:InterventionTypeScheme`, and `erct:OutcomeDomainScheme`. Facets are now modelled to use controlled IRIs for stable filtering and reuse, while acknowledging that legacy records may still contain literal values.

5.6.4 User Study

The study employed a within-subjects, counterbalanced design in which each participant completed the same task set under all three interface conditions (C0–C2; see citation visibility manipulation above). Condition order was counterbalanced to mitigate learning and fatigue effects. The procedure was designed to examine how mediation and provenance visibility influence retrieval, interpretation, synthesis, and verification of international development evidence, as well as workload, trust, and reliance behaviour. Ethics for this study was approved by the Research Ethics Committee at Trinity College Dublin (REF 4740).

5.6.4.1 Participants

Twenty-one participants completed the study ($n=21$). Gender was balanced (10 male, 10 female). Ages were primarily 25–34 (73.7%), with 18–24 (21.1%) and 65+ (5.3%). Education levels were predominantly postgraduate: Master’s degree (52.6%), Doctoral degree (21.1%), Bachelor’s degree (15.8%), and Postgraduate diploma/certificate (10.5%). Participants reported moderate-to-high familiarity with LLMs (scores clustered at 4/5; 57.9%), and 42.1% reported awareness of development evidence platforms. 84.1% reported reading international development reports “Somewhat frequently” or higher. Participants were compensated for their time with a €20 gift voucher.

5.6.4.2 Study Procedure

Each study session was conducted synchronously via video conferencing and lasted approximately 1 hour. Participants completed the study remotely on their own devices,

using a desktop web browser. All study flow, instructions, questionnaires, and task transitions were delivered through a Qualtrics survey, while task execution took place in the InterDev system via embedded links. All study documentation is available on the thesis GitHub repository in the iteration 3 folder²¹.

After providing informed consent, participants first completed a short pre-study questionnaire capturing demographic information, prior exposure to international development research, familiarity with large language models, and related background variables. These measures were used to contextualize individual differences in performance and subjective experience.

Participants then completed three evaluation blocks, one for each interface condition. The order of interface conditions was counterbalanced across participants using a Latin Square design, ensuring that each condition appeared equally often in the first, second, and third block positions. Task order was independently balanced such that each participant completed six experimental tasks in total (two per block), covering all task themes without repetition. The task set in Table 15 was designed to approximate the kinds of evidence-use episodes that occur in development research and policy support, while remaining feasible within a one-hour session. Each domain theme (e.g., agriculture, health systems, social protection, education, gender/livelihoods) was chosen to reflect common intervention areas in impact-evaluation evidence and to ensure that participants encountered variation in interventions, outcome types, and contextual descriptors. Tasks were framed as realistic policy-oriented prompts rather than abstract lookup questions, because the central focus of Iteration 3 is not retrieval alone but higher-order evidence use: comparing studies, synthesising implications, and justifying recommendations under uncertainty (R8).

Each evaluation block followed the same structured sequence. First, participants were given a brief facilitator-led demonstration of the assigned interface, highlighting only the core interaction mechanisms relevant to that condition. This was followed by a short tutorial task designed to allow participants to familiarize themselves with the interface in a low-stakes setting; tutorial responses were not included in the analysis. Within each condition block, a short tutorial task was included to reduce interface novelty effects and

²¹ <https://github.com/mmurtaghw/InterDevResearch>

to stabilise performance measures before the two scored tasks. The main tasks were constructed to require (i) retrieval of multiple evaluations, (ii) cross-study comparison, and (iii) an explicit written synthesis, so that accuracy and reasoning quality could be assessed consistently across conditions. Prompts were interface-agnostic by design: they can be solved via baseline browsing (C0) or via conversational assistance (C1/C2), but they impose the same underlying requirements (e.g., “use at least two evaluations”) so that differences in workload, confidence calibration, and verification behaviour can be attributed to mediation and provenance visibility rather than to task feasibility. The inclusion of multi-evaluation prompts also creates a natural opportunity for citation checking in C2, enabling analysis of whether visible provenance changes verification strategies (R5–R6) and whether any workload reduction occurs without degrading judgement quality (R7). The full study protocol is available in Appendix A. The protocol and all associated study tasks and documentation are available in the thesis GitHub repository under the iteration 3 folder²².

Task Type	Task Prompt	Domain
Tutorial	Compare two or more impact evaluations where farmers faced climate or weather shocks; note the support provided and one key outcome for each.	Agriculture / Climate
Main Task	Identify agricultural interventions that improve smallholder outcomes in Sub-Saharan Africa, comparing at least two impact evaluations and summarizing outcome differences.	Agriculture
Main Task	Identify and recommend a program to improve child nutrition and health in resource-constrained settings using evidence from at least two randomized impact evaluations.	Health / Nutrition
Tutorial	Summarize two or more evaluations aimed at slowing disease spread, describing what changed and the outcomes.	Health
Main Task	Advise a health ministry on improving service quality and motivation among health workers based on evidence from at least two impact evaluations.	Health Systems
Main Task	Describe three or more social protection programs targeting food security and synthesize what they collectively suggest for policy.	Social Protection
Tutorial	Identify two digital learning interventions, specifying the tool used and reported outcomes.	Education

²² <https://github.com/mmurtaghw/InterDevResearch>

Task Type	Task Prompt	Domain
Main Task	Propose a program to support female microentrepreneurs based on intervention evidence and reported outcomes.	Gender / Livelihoods
Main Task	Recommend policy options to improve learning outcomes in low-income regions.	Education

Table 15: Tasks completed by participants during the evaluation

Participants were then notified that the formal evaluation phase was beginning and proceeded to complete two experimental tasks within the same policy domain theme. All tasks outlined in Table 15 were interface-agnostic and could be solved using any of the provided tools, requiring participants to retrieve, interpret, compare, and synthesize evidence from curated collections within InterDev. Participants completed the tasks independently, without assistance from the facilitator, who was present only to provide the scripted demonstration of each condition and to clarify procedural questions before the task began. Examples of such procedural questions were how to use Qualtrics and where to fill in answers on the form. Upon completing the two tasks in a block, participants filled out a post-condition questionnaire. This included the NASA Task Load Index (NASA-TLX) to assess perceived workload, along with additional Likert-scale items measuring trust, perceived accuracy, and verification behaviour specific to the interface just used. It also included the Trust in Automation (TIA) scale, a 12 item Likert-scale used to measure trust in automated systems (Jian et al., 2000). This block-level assessment was repeated after each interface condition.

After completing all three interface blocks, participants completed a Post-Study System Usability Questionnaire (PSSUQ) assessing overall system usability across conditions. Finally, participants took part in a short post-study interview guided by structured prompts, which can be seen on Table 16. This discussion focused on participants' comparative preferences, perceived trustworthiness of each interface, task-interface fit, moments of high or low confidence, and how the presence or absence of surfaced citations influenced verification and reliance behaviour.

ID	Interview question prompt
Q1	Of the three interfaces you used (Baseline, Chat, Chat + KG Citations), which did you prefer overall, and why?
Q2	Which interface felt most supportive when you needed to compare or combine findings across multiple evaluations or sources?
Q3	During the tasks, were there particular moments where you felt especially confident in the answer you were producing? Please describe.
Q4	How did the presence or absence of AI support (and citation visibility) influence how much you trusted the information presented?
Q5	Thinking about cognitive effort, which interface felt easiest to work with, and why?
Q6	Is there anything you wanted the system to do that none of the interfaces currently supported?

Table 16: Participant prompts used for the post-evaluation interview.

5.6.5 Measures

The following measures were introduced specifically for iteration 3:

Task Performance. Each experimental task response was scored by the author of this thesis using a structured rubric tailored to the task prompts. Scores ranged from 0 to 4 based on evidence use, completeness, and synthesis quality (see rubric). These scores served as the primary task performance outcome. To reduce rater bias and establish reliability, a second independent rater scored a randomly selected subset of 30% of responses using the same rubric. Using Cohen’s κ , an inter-rater agreement of 0.72 was determined.

Task Efficiency. Completion time was derived from the Qualtrics page-submit timestamp for each task ($T^* \text{ Time_Page Submit}$). Task-time data in usability evaluations are typically positively skewed, so analyses commonly apply a log transformation to stabilise variance and reduce undue influence of extreme values (Sauro & Lewis, 2010). To reduce skew, completion times were therefore analysed on the log scale.

Subjective Workload (NASA–TLX). After each condition, participants completed the NASA–TLX subscales (Mental, Physical, Temporal, Performance, Effort, Frustration) (Hart & Staveland, 1988). In the analysis, both overall workload, computed as the mean of the six subscales, and each subscale-specific scores were reported.

Trust in Automation. Trust was measured using the Trust in Automation (TIA) scale, consisting of 12 items per condition block (Jian et al., 2000). The items assess whether the system is perceived as deceptive, underhanded, and whether the user feels suspicious or wary, and whether its actions might be harmful, alongside positive judgments including confidence, security, integrity, dependability, reliability, overall trust, and familiarity with the system. Items were averaged to form a per-condition trust score.

Confidence and Calibration. After each task, participants reported their confidence on a 0–100 scale. These ratings were rescaled to the unit interval and interpreted as subjective probabilities of being correct (Ma et al., 2024b). Calibration was assessed using a Brier-style squared-error score, computed as the mean squared difference between reported confidence and task performance. The Brier score is a standard proper scoring rule for evaluating probabilistic judgments and is widely used in calibration research; when performance is binary, this formulation reduces to the canonical Brier score (Brier, 1950). Because task responses were graded using a 0–4 rubric, performance was operationalised by normalising the rubric score, yielding a continuous measure of correctness. This preserves the same squared-error interpretation while allowing confidence–performance alignment to be evaluated under partial-credit scoring.

5.6.6 Results and Analysis

Linear mixed-effects models were fitted to all quantitative outcomes, using participant random intercepts to account for repeated measures. For task-level outcomes (rubric score, log-transformed completion time, confidence, and Brier score), models included task fixed effects to control for task difficulty. Condition (C0, C1, C2) was modelled as a fixed effect, with C0 treated as the baseline. For condition-level outcomes (NASA–TLX and TIA), analyses were conducted on per-participant means by condition. This modelling approach follows standard practice in quantitative HCI for repeated-measures user studies, where mixed-effects models are recommended to account for within-participant dependence and other grouping structures that are common in interaction data (Robertson & Kaptein, 2016).

All models were estimated using maximum likelihood. Reported coefficients represent differences relative to the baseline condition (C0), with p -values derived from z -tests. For

presentation, condition means are reported with standard errors and annotate figures with significance markers.

5.6.7 Quantitative Results

Quantitative outcomes are reported from a within-subjects study ($n = 21$), analysed using linear mixed-effects models with participant random intercepts. For task-level outcomes (scores, completion time, confidence, and calibration), models included task as a fixed effect. Completion time was log-transformed prior to analysis. The anonymised raw data of this evaluation is available in the thesis GitHub repository under the iteration 3 folder²³. No significant differences were observed between the different demographic groups sampled in Section 5.6.4.1.

5.6.7.1 Task Performance and Efficiency

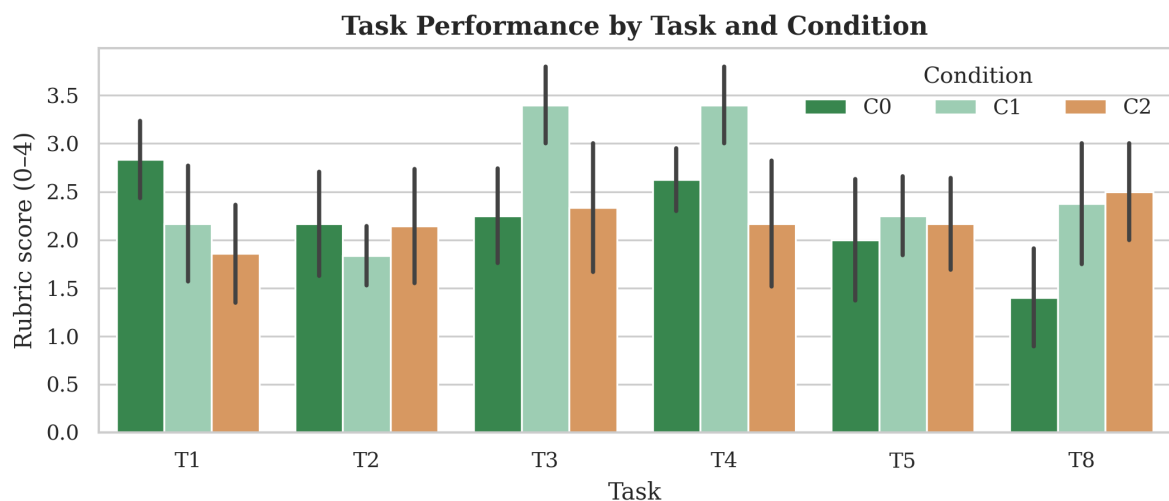


Figure 23: Task performance by task and condition. C0 indicates the baseline condition, C1 LLM-mediated exploration (no provenance cues) and C2 LLM-mediated exploration with provenance cues.

Task Performance

Linear mixed-effects models on rubric scores revealed no reliable effects of condition after controlling for task difficulty through task fixed effects. Figure 23 shows that, relative to the baseline condition (C0), condition C1 showed a small positive trend ($\beta = +0.320$, $p = 0.158$), whereas C2 did not differ from C0 ($\beta = -0.026$, $p = 0.909$). Task effects were also nonsignificant (all $p > 0.16$). Overall, task performance appears to be dominated by task difficulty rather than interface condition.

²³ <https://github.com/mmurtaghw/InterDevResearch>

Completion Time

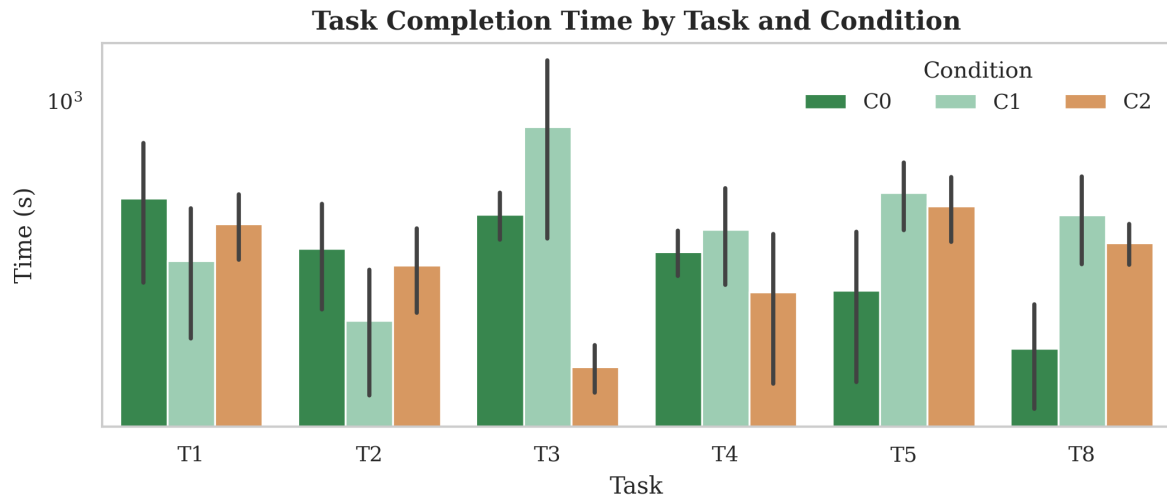


Figure 24: Task Completion Time by Task and Condition. Time is log transformed to better fit the linear model. C0 indicates the baseline condition, C1 LLM-mediated exploration (no provenance cues) and C2 LLM-mediated exploration with provenance cues.

Models on log-transformed completion time similarly showed no reliable differences between conditions (C1 vs. C0: $\beta = +0.071$, $p = 0.578$; C2 vs. C0: $\beta = -0.079$, $p = 0.531$). Figure 24 displays a marginal task effect indicated that task T2 was completed faster than task T1 ($\beta = -0.322$, $p = 0.071$). These results suggest that interface condition did not meaningfully affect efficiency in this sample.

5.6.7.2 Workload (NASA-TLX)

Overall Workload

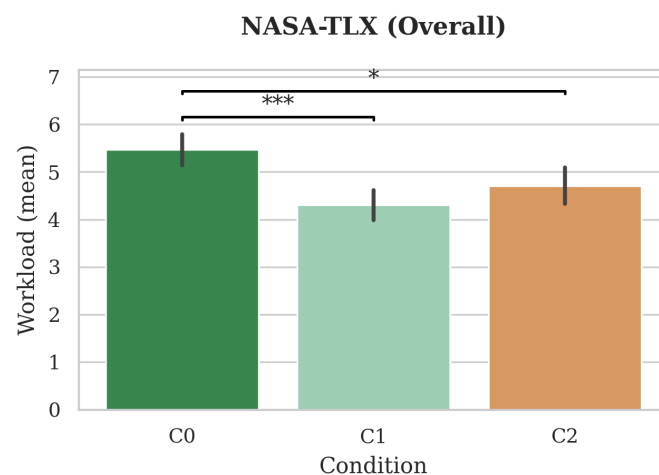


Figure 25: Overall Aggregated NASA-TLX workload score. C0 indicates the baseline condition, C1 LLM-mediated exploration (no provenance cues) and C2 LLM-mediated exploration with provenance cues (* $p < .05$, ** $p < .01$, *** $p < .001$).

Overall workload decreased under both chat conditions (Figure 25). Compared to baseline (C0), condition C1 significantly reduced NASA–TLX scores ($\beta = -1.163$, $p = 0.001$), and condition C2 also reduced workload ($\beta = -0.754$, $p = 0.032$). Mean workload scores were C0 = 5.46, C1 = 4.30, and C2 = 4.71.

Subscale Analysis

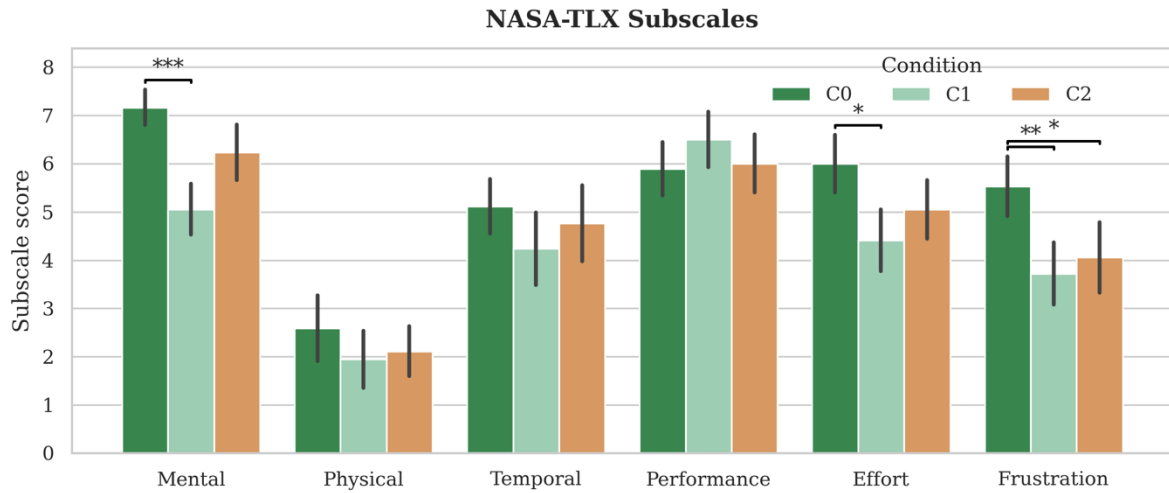


Figure 26: NASA-TLX workload subscales, C0 indicates the baseline condition, C1 LLM-mediated exploration (no provenance cues) and C2 LLM-mediated exploration with provenance cues (* $p < .05$, ** $p < .01$, *** $p < .001$).

Figure 26 shows that workload reductions were primarily concentrated in the Mental, Effort, and Frustration subscales (TLX item order: Mental, Physical, Temporal, Performance, Effort, Frustration). Mental Demand was significantly reduced in C1 relative to baseline ($\beta = -2.140$, $p < 0.001$), with a smaller and nonsignificant reduction observed in C2 ($\beta = -0.991$, $p = 0.114$). A similar pattern was observed for Effort, which decreased in C1 ($\beta = -1.558$, $p = 0.034$) but not reliably in C2 ($\beta = -0.965$, $p = 0.174$). Frustration was reduced under both chat conditions, with significant effects in C1 ($\beta = -1.465$, $p = 0.006$) and C2 ($\beta = -1.124$, $p = 0.033$). Physical Demand, Temporal Demand, and Performance showed no reliable differences across conditions (all $p > 0.14$). Together, these patterns suggest that chat interfaces reduce perceived mental demand and effort, with the strongest reductions observed in condition C1.

5.6.7.3 Confidence and Calibration

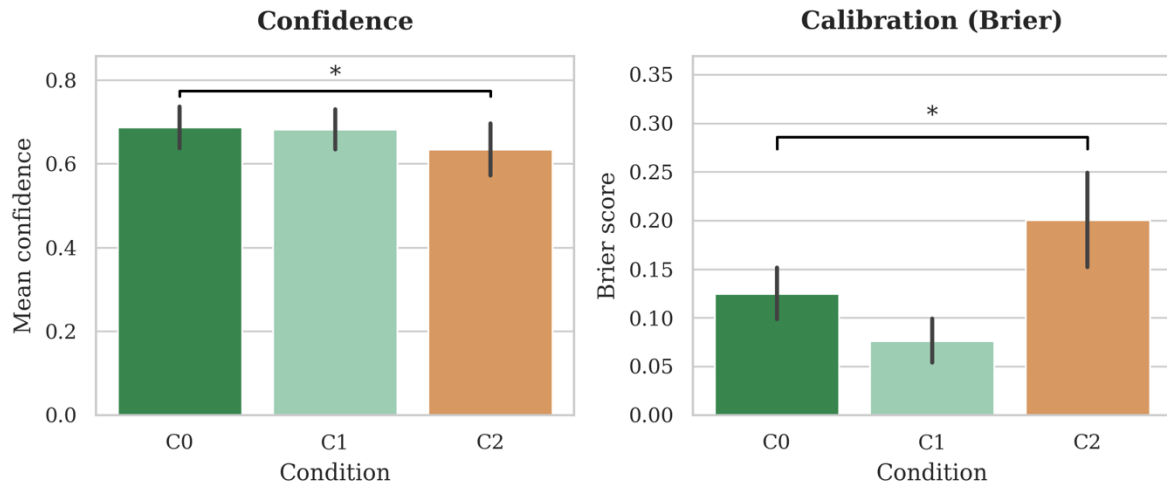


Figure 27: Self-Reported Confidence and Brier confidence calibration measures. C0 indicates the baseline condition, C1 LLM-mediated exploration (no provenance cues) and C2 LLM-mediated exploration with provenance cues (* $p < .05$, ** $p < .01$, *** $p < .001$).

Self Reported Confidence

Mixed-effects models on logit-transformed confidence indicated that condition C2 reduced confidence relative to baseline ($\beta = -0.461$, $p = 0.034$), as per Figure 27. Condition C1 did not differ from C0 ($\beta = +0.002$, $p = 0.994$). Mean confidence values were C0 = 0.687, C1 = 0.682, and C2 = 0.634.

Confidence Calibration

Calibration was assessed using Brier scores, with normalized rubric scores serving as the correctness proxy (Figure 27). Calibration worsened in condition C2 relative to C0 ($\beta = +0.075$, $p = 0.043$). Condition C1 showed a nonsignificant trend toward improved calibration ($\beta = -0.054$, $p = 0.152$). Task T8 exhibited significantly worse calibration than task T1 ($\beta = +0.141$, $p = 0.007$). Mean Brier scores were C0 = 0.125, C1 = 0.076, and C2 = 0.200.

5.6.7.4 Trust in Automation

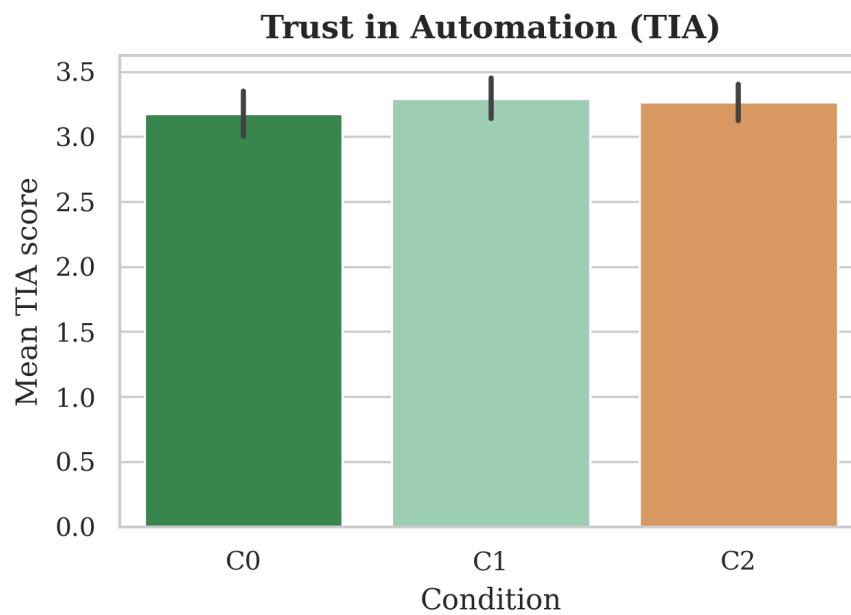


Figure 28: Jian et al Trust in Automation measure. C0 indicates the baseline condition, C1 LLM-mediated exploration (no provenance cues) and C2 LLM-mediated exploration with provenance cues.

Trust in automation, measured using the Jian et al. Trust in Automation (TIA) scale, did not differ reliably across conditions. Figure 28 shows that relative to baseline, C1 showed $\beta = +0.118$ ($p = 0.406$), and C2 showed $\beta = +0.088$ ($p = 0.538$). These results indicate that despite changes in workload and confidence, reported trust remained stable across interface conditions

5.6.7.5 PSSUQ

Post-study usability (PSSUQ), as indicated by Figure 29, indicated moderate perceived usability overall. The mean overall score was 3.27/7, suggesting that participants' experience was mixed rather than strongly positive.

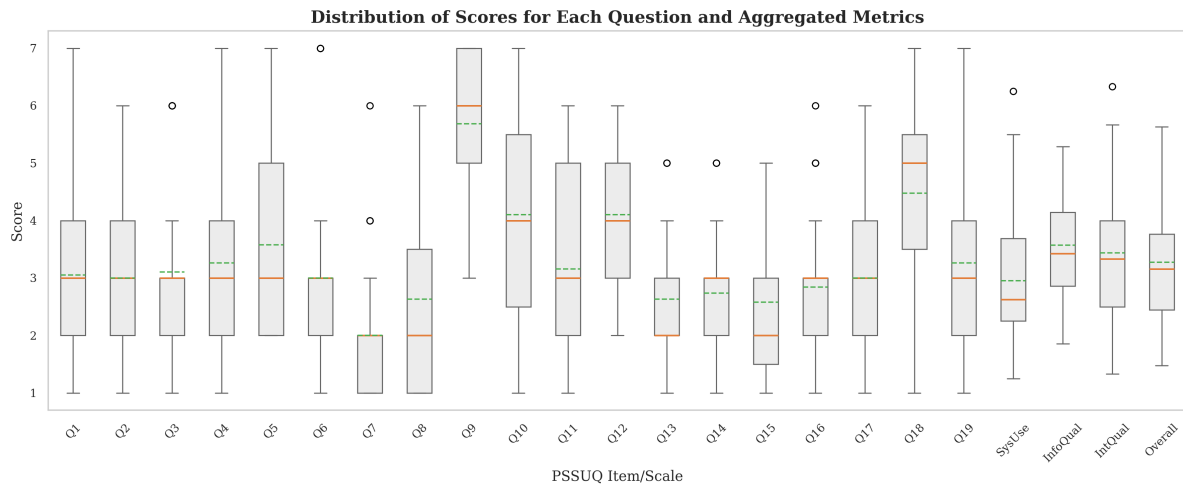


Figure 29: Distribution of PSSUQ item scores and aggregated metrics (SysUse, InfoQual, IntQual, Overall). Lower scores are better.

Subscale patterns showed that System Usefulness was weakest (2.95/7 inverted), while Information Quality was relatively better (3.57/7) and Interface Quality was intermediate (3.44/7). Item-level distributions in Figure 29 show substantial spread across participants, with several items clustering toward lower usability ratings, especially within the System Usefulness set.

Compared to Iteration 1 and Iteration 2, Iteration 3 shows a general shift toward less favourable PSSUQ ratings across items. This change should be interpreted in light of Iteration 3’s more complex, synthesis-oriented tasks and added interaction mechanisms, which increase evaluative and interpretive demands beyond earlier task sets. System Usefulness was weakest (2.95/7), while Information Quality was relatively better (3.57/7) and Interface Quality was intermediate (3.44/7). Distributions remain wide, and as in earlier iterations, items related to feedback and error handling (notably Q9) show greater variability, indicating persistent issues with recovery support and message clarity.

5.6.7.6 Summary of Quantitative Patterns Findings

Across measures, chat interfaces reduced perceived workload, with the strongest effects observed in condition C1. However, neither task performance nor completion time improved reliably once task effects were controlled. The LLM-mediated exploration with provenance cues condition (C2) was associated with lower confidence and worse calibration relative to baseline, despite reduced workload, while trust ratings remained

unchanged. Taken together, these findings suggest that adding citations did not increase performance or trust and may instead have introduced a caution signal in the form of lower confidence, without corresponding improvements in calibration.

5.6.8 Qualitative Results

End-of-session interviews were analysed using Braun and Clarke’s thematic analysis described in Section 5.3.3.2 (Braun & Clarke, 2006), focusing on participants’ interface preferences, verification behaviours, confidence calibration, and perceived usability frictions. Five dominant themes emerged. An example of processed data from these interviews is available on the thesis GitHub repository²⁴. In brief, the discussion that follows is structured around five themes identified in the qualitative analysis. These are: the efficiency–assurance trade-off; sources as trust anchors with conditional trust; chat as a facilitator of comparison but with limited perceived synthesis and search and navigation friction. Table 17 shows the incidence of references within the text for each of the identified themes. Figure 30 shows the distribution of code frequencies across participants, detailing the codes used as part of the analysis.

Themes	Findings	References	Scope
Efficiency-assurance trade-off	Participants described chat as faster and lower effort, but often with lower confidence; chat plus sources was treated as the best compromise between speed and assurance.	157	Workflow
Sources as trust anchors with conditional trust	Visible sources increased perceived credibility, but participants still emphasized verification and double-checking rather than accepting outputs at face value.	129	Trust
Search and navigation friction	Rigid search behaviour and interface friction (filters, scrolling, interaction controls) constrained effective use; participants requested richer query and filtering options.	71	UI/Search
Chat facilitates comparison but synthesis is limited	Chat supported comparison across papers and reduced manual cross-referencing, but some participants reported that outputs remained descriptive and not fully integrative.	59	Synthesis

Table 17: Summary of qualitative themes, key findings, reference counts, and analytic scope.

²⁴ <https://github.com/mmurtaghw/InterDevResearch>

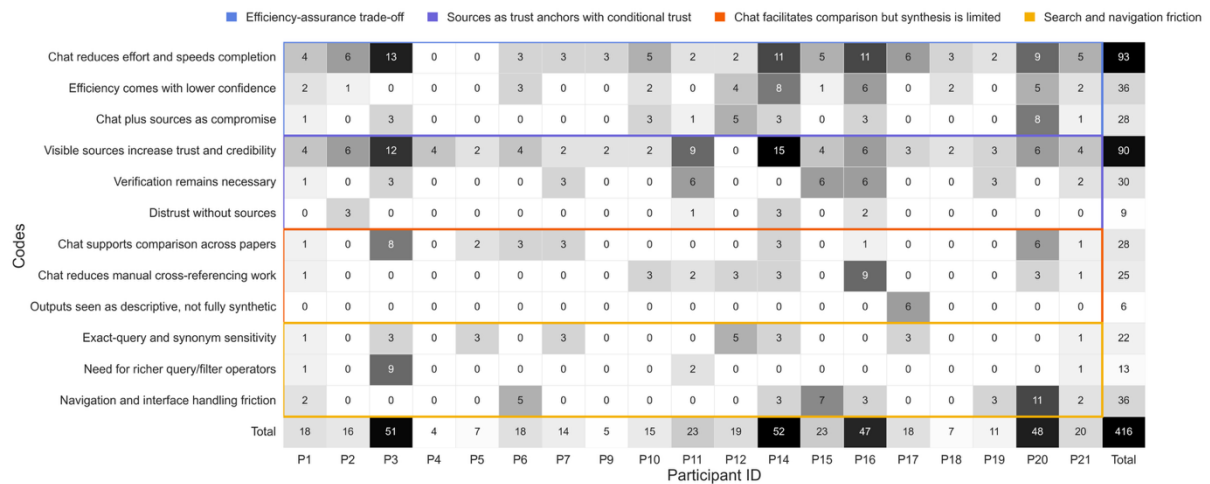


Figure 30: Distribution of qualitative code frequencies across participants, organised by overarching theme. Darker cells indicate higher code counts, and marginal totals summarise the number of references per code and per participant.

Efficiency–assurance trade-off. Participants consistently framed their interface preferences as a trade-off between efficiency and assurance. Chat-based interaction was widely described as reducing effort and accelerating task completion, yet this efficiency gain was often perceived to come “at some price on the confidence” (P12). Several participants noted that chat without sources was “the easiest” option, but one participant emphasized that this ease was accompanied by feeling “less confident” (P16). Others described chat with sources as a more acceptable compromise, with one participant stating that “chat plus sources was a better one” (P10). This theme mirrors the quantitative reductions in perceived workload under chat conditions, while highlighting that efficiency alone was insufficient to establish confidence.

Sources as trust anchors, with conditional trust. Across interviews, participants treated sources as an important anchor for credibility, yet trust in system outputs remained explicitly conditional. Some participants reported higher trust when they could see that the system was “only referring to the papers that I had added to the collection” (P11), while others expressed a more categorical stance, stating that they “don’t trust AI unless I can see sources” (P14). However, even when sources were available, participants emphasized ongoing scepticism, describing their engagement as taking results “with a pinch of salt” and emphasizing a continued need to “double-check” (P15). These accounts suggest that citations functioned as a prerequisite for trust rather than a substitute for verification.

Chat facilitates comparison, but synthesis is perceived as limited. Participants frequently highlighted the value of chat for comparing across multiple papers and reducing the effort of cross-referencing. One participant described the task of reviewing different papers as “so much easier with a chat” (P14), particularly relative to manual browsing. At the same time, some participants perceived the resulting outputs as insufficiently integrative. For example, one participant observed that the system would “describe the contributions” of individual papers but “wouldn’t necessarily synthesize them” into a coherent whole (P17). This tension suggests that while chat supported surface-level comparison, it did not always meet participants’ expectations for higher-level synthesis.

Search and navigation friction constrain effective use. Usability friction related to search and navigation was a recurring concern and shaped participants’ overall experience. Several participants reported frustration with rigid search behaviour, noting that the system required “an exact search query,” such as using “low income instead of low resource” (P12). Others expressed a desire for more expressive queries, wishing they could search for “multiple things... like food plus policy” (P11). Additional friction arose from interaction design issues, including lost saved searches (P06) and cumbersome filtering interfaces that required excessive scrolling, described as having to go “down, down, down” (P15). These limitations constrained effective use even when chat functionality was available.

5.6.8.1 Summary of Qualitative Patterns Findings

Participants described chat-based interaction as a valuable mechanism for reducing effort and supporting comparison across documents, but one that introduced trade-offs related to confidence, trust, and depth of synthesis. Sources played a central role in shaping perceived credibility, yet trust remained conditional and closely tied to a user’s own verification practices. Persistent search and navigation frictions, together with an initial learning curve, further influenced how participants engaged with the system and the extent to which they relied on its outputs.

5.6.9 Limitations

Iteration 3 strengthens internal validity by using a within-subjects, counterbalanced design, but several limitations constrain the scope and interpretation of the findings. The sample size was modest ($n = 21$), which limits power to detect small condition effects and increases

sensitivity to individual differences in domain familiarity, LLM experience, and personal verification habits. In addition, the session length and remote setting introduced natural constraints on how deeply participants could engage in checking sources and reconciling findings, meaning observed verification behaviour may reflect time-bounded study conditions rather than how users would behave in higher-stakes or extended synthesis work.

Although the interface manipulation was designed to vary only mediation and citation visibility, the conditions are not psychologically equivalent in cognitive demands. Chat introduces an additional interaction channel and may change task strategy, while citations can add reading and cross-referencing overhead. As a result, differences in workload, confidence, and calibration may partly reflect shifts in strategy and attention allocation rather than changes in underlying capability or evidence comprehension. Relatedly, the “bounded collection” constraint improves grounding but may reduce ecological validity, since real evidence work often involves iteratively expanding and revising the evidence set beyond an initial working collection.

Several measures also impose interpretation limits. Task performance was scored using a 0–4 rubric, which is necessary for synthesis tasks but introduces subjectivity and potential ceiling and floor effects; while inter-rater agreement was acceptable, the scoring remains an imperfect proxy for evidence reasoning quality. Confidence calibration was operationalised using a Brier-style score on normalised rubric values, which enables continuous scoring but assumes that rubric points map linearly onto “correctness” in a way that may not fully capture partial evidence use, justificatory quality, or uncertainty-sensitive reasoning. NASA–TLX and trust scales are self-reported and can be influenced by interface fluency, user expectations about LLMs, and social desirability effects, particularly in conditions where assistance is salient.

5.6.10 Iteration 3: Conclusions

Across quantitative outcomes and qualitative measures, a consistent pattern emerges. There are indications that chat-based interaction reliably reduces perceived workload, but does not translate into measurable gains in task performance or efficiency, and the addition of citations reshapes users’ engagement in ways that affect confidence without increasing trust. Taken together, the results suggest that chat primarily functions as a mechanism for

reducing interactional effort, while epistemic effort, including verification, synthesis, and judgment, remains largely unchanged. Citations, rather than serving as trust-enhancing cues, appear to introduce a form of epistemic caution that alters a user's subjective confidence and calibration without improving objective outcomes.

Chat reduced workload without improving performance

Empirically, the chat-based conditions reduced perceived workload relative to the baseline, especially on the NASA-TLX mental demand, effort, and frustration subscales. Participants also consistently described the chat interfaces as easier, faster-feeling, and less taxing to use.

Interpretively, this suggests that chat reduced the interactional cost of retrieval and comparison. It appeared to make it easier for participants to access information, reformulate queries, and move through the evidence space. However, these gains did not extend to reliable improvements in task performance or efficiency. One plausible explanation, supported by the interviews, is that effort was not removed so much as redistributed: participants described continuing to spend time on verification, reformulation, and interpretation. In other words, the interface reduced the cost of access, but not the epistemic work required to judge and integrate evidence.

As a design implication, this suggests that conversational assistance alone is unlikely to improve outcomes in tasks dominated by synthesis and judgement. If the goal is to improve task performance rather than only reduce felt effort, the system likely needs stronger support for comparing, reconciling, and integrating evidence, not just easier retrieval.

Citations appeared to prompt caution rather than increase trust

The most striking divergence between conditions emerged in the LLM-mediated exploration with provenance cues interface. While citations were seen to reduce perceived workload relative to baseline, they were associated with lower self-reported confidence and worse confidence calibration, without any corresponding increase in trust in automation. At first glance, this pattern appears paradoxical, as added transparency might be expected to increase confidence or trust. The qualitative data seems to resolve this tension.

Participants consistently treated citations as signals to scrutinize rather than as endorsements of correctness. Seeing sources prompted users to slow down, question outputs, and reassess their own certainty. Rather than reassuring participants, citations often highlighted uncertainty by revealing gaps, inconsistencies, or the need for additional checking. In this sense, citations functioned as a caution signal that foregrounded epistemic responsibility rather than offloading it to the system.

This interpretation helps reconcile three otherwise puzzling findings. These findings are reduced confidence, degraded calibration, and unchanged trust. It would seem that trust ratings remained stable because participants' underlying stance toward the system did not shift. Trust was already conditional and bounded. Confidence, by contrast, reflected participants' assessment of their own answers, which decreased when uncertainty became more salient. The worsening of calibration in the citation condition suggests that increased awareness of uncertainty does not necessarily improve users' ability to align confidence with correctness, particularly when verification is effortful and synthesis support is limited.

As a design implication, this suggests that conversational assistance alone is unlikely to improve outcomes in tasks dominated by synthesis and judgement. If the goal is to improve task performance rather than only reduce felt effort, the system likely needs stronger support for comparing, reconciling, and integrating evidence, not just easier retrieval.

Conditional trust and active verification as the dominant mode of use

Conditional trust and active verification emerged from the post-study interview section, specifically in response to question 4. Across interviews, participants articulated a consistent model of interaction. The system was useful, but never authoritative. Trust was framed as conditional and contingent on visible sources, personal judgment, and external verification. Even participants who stated that they do not trust AI without sources emphasized that sources were a prerequisite for engagement rather than a substitute for checking. During interviews, 17 participants explicitly referenced sources and citations when discussing trust, and 8 explicitly described verification behaviours, suggesting that participants treated the system as useful but not authoritative.

This pattern helps explain why trust scores did not vary across conditions despite meaningful changes in workload and confidence. Participants did not move toward greater

reliance or delegation under chat conditions. Instead, they maintained an active role in evaluating and validating information. The system was treated as an assistant for sensemaking rather than an autonomous decision-maker. From this perspective, stable trust scores are not a failure of the design but an accurate reflection of sustained human oversight.

These findings align with prior accounts of human–AI interaction in which users aim for appropriate reliance rather than treating AI outputs as authoritative. Trust functions as a regulator of reliance under uncertainty, and stable trust can reflect sustained human oversight rather than a design failure (Lee & See, 2004). Importantly, they also suggest that increased transparency and provenance cues do not automatically translate into higher trust. Human-centred explainable AI research argues that transparency alone is insufficient and that trust is shaped by task context, user goals, and the practices users employ to assess reliability (Zerilli et al., 2022). In LLM settings specifically, citation mechanisms can shape both trust and users’ willingness to verify claims by checking sources, reinforcing the view that trust is embedded in broader verification and domain-judgment practices rather than produced by transparency in isolation (Ding et al., 2025).

Synthesis, Not Retrieval, as the Primary Bottleneck

While participants widely praised chat for facilitating comparison across documents and reducing the burden of cross-referencing, several noted that system outputs often stopped short of true synthesis. Responses were described as enumerating contributions or summarizing individual papers rather than integrating findings into a coherent, higher-level account. This perceived lack of synthesis helps explain why chat did not yield performance gains despite lowering workload.

The system effectively supported access to information but offered limited scaffolding for integration. As a result, users were left to perform the most cognitively demanding aspects of the task themselves. These aspects included reconciling claims, weighing evidence, and constructing overarching interpretations. This limitation likely constrained both task performance and calibration, particularly in the citation condition where verification demands were more salient.

From a design perspective, these findings suggest that retrieval- and comparison-oriented assistance may deliver diminishing returns once basic access barriers are removed. Supporting synthesis through clearer claim to evidence alignment, contrastive summaries, or explicit reasoning structures may be necessary to translate reduced effort into improved outcomes.

5.6.11 Ontology changes as a result of iteration 3

Iteration 3 showed that interaction-level transparency changes how users judge and verify system outputs, even when task performance is unchanged. Participants consistently treated sources as a prerequisite for credibility, but citations also acted as a caution signal that lowered confidence and did not reliably improve calibration. Qualitative feedback further indicated that the primary bottleneck had shifted from retrieval to synthesis: chat reduced mental demand and effort, yet users still needed to compare, reconcile, and integrate findings across multiple impact evaluations. These results motivated ontology changes that better support verification and synthesis in ways that are durable beyond a specific interface, by improving lexical alignment for discovery and by introducing a more granular unit of evidence that can be referenced and compared.

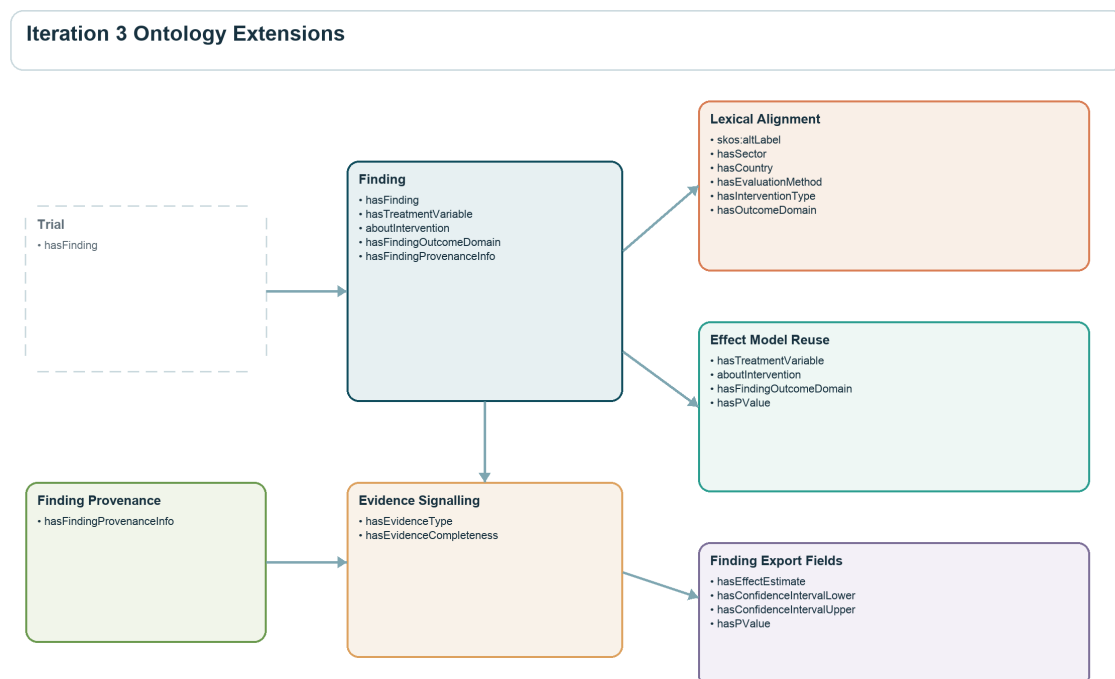


Figure 31: Iteration 3 ERCT ontology, showing the finding-level evidence model, finding provenance, evidence signalling properties, lexical alignment properties, and export/effect-reuse properties added to support verification and synthesis.

In response, iteration 3 extended the ontology to strengthen evidence reuse and cross-impact evaluation verification. First, SKOS coverage was expanded across controlled facet schemes to better match user phrasing during search and filtering, adding `skos:altLabel` variants while preserving canonical facet properties (`erct:hasSector`, `erct:hasCountry`, `erct:hasEvaluationMethod`, `erct:hasInterventionType`, `erct:hasOutcomeDomain`). Second, a finding-level evidence model was introduced via `erct:Finding` and `erct:hasFinding` to support citation and comparison at the level users actually reason over, rather than only at the impact evaluation level. Findings reuses existing effect nodes through `erct:hasTreatmentVariable` (`erct:Finding` → `erct:TreatmentVar`) and can be linked to interventions (`erct:aboutIntervention`) and outcome domains (`erct:hasFindingOutcomeDomain`), with export-oriented numeric fields added (`erct:hasEffectEstimate`, `erct:hasConfidenceIntervalLower`, `erct:hasConfidenceIntervalUpper`) while continuing to reuse `erct:hasPValue` on `erct:TreatmentVar`. Finally, controlled indicators for evidence type and completeness were added using `erct:hasEvidenceType` and `erct:hasEvidenceCompleteness` over dedicated SKOS schemes, supporting practical synthesis and quality signalling in exports.

5.7 FOOPS Review of Ontology

Throughout the user evaluations described in this chapter, the ERCT ontology has been changed and extended in response to the needs of the interaction workflow and user feedback. To ensure that these iterative changes also remained aligned with recognised ontology quality and FAIR-oriented publishing expectations, the final ontology version was assessed using FOOPS!²⁵, a tool that evaluates ontologies against a set of checks spanning findability, reuse-oriented metadata, and vocabulary quality. The goal of this review was not to determine whether ERCT is conceptually complete for all development evidence, but to verify that the ontology artefact is consistently documented, internally usable, and ready for long-term reuse in tooling and data pipelines.

²⁵ https://github.com/oeg-upm/fair_ontologies

The FOOPS review compared the third iteration baseline run against a final remediated run where potential FAIR issues were identified and addressed. The baseline achieved an overall score of 0.90, with 13 checks passing and 2 failing. After targeted ontology-header remediation, the score increased to 0.9333, with 14 checks passing and 1 failing. The only check that changed status was OM3, which passed following the addition of standard metadata annotations and vocabulary prefixes in the ontology header, including DOI identifiers, status information, and a logo link. The remaining failure, PURL1, persisted across both runs and reflects an infrastructural identifier requirement rather than a modelling flaw. Specifically, the ontology IRI still uses a non-persistent pattern, which FOOPS does not recognise as a persistent identifier scheme. As a result, the FOOPS outcomes indicate that the adjusted post-iteration 3 ERCT meets the tool’s documentation, vocabulary, and reuse checks, while the outstanding issue is the adoption of a persistent, long-term resolvable ontology IRI²⁶. Details of the assessment results from the various iterations of the ontology can be found in Table 18.

²⁶ The final WIDOCO documented ontology is available in the thesis GitHub (<https://github.com/mmurtaghw/InterDevResearch>)

FOOPS check	Category	V5 baseline	V5 final	Change?	
Overall	Summary	0.9000	0.9333	Improved	+0.0333 overall.
Pass count / total	Summary	13 / 15	14 / 15	Improved	One additional check passed.
Failure count / total	Summary	2 / 15	1 / 15	Improved	One remaining failure.
PURL1	Findable	Failure	Failure	No	Persistent identifier policy for the ontology IRI.
RDF1	Interoperable	Pass	Pass	No	RDF-level compliance check passed.
OM1	Reusable	Pass	Pass	No	Core ontology metadata expectations met.
OM2	Reusable	Pass	Pass	No	Optional contributor metadata may still be incomplete. Not a failure.
OM3	Reusable	Failure	Pass	Yes	Fixed by adding header metadata and required prefixes, including DOI identifiers, status, and a logo link.
OM4_1	Reusable	Pass	Pass	No	License presence check passed.
OM4_2	Reusable	Pass	Pass	No	License resolvable check passed.
OM5_1	Reusable	Pass	Pass	No	
OM5_2	Reusable	Pass	Pass	No	Additional metadata reuse check passed.
FIND1	Findable	Pass	Pass	No	Findability related check (non-PURL) passed.
VOC1	Interoperable	Pass	Pass	No	Vocabulary usage check passed.
VOC2	Interoperable	Pass	Pass	No	Vocabulary reuse/structure check passed.
VOC3	Interoperable	Pass	Pass	No	Labels complete enough to pass.
VOC4	Interoperable	Pass	Pass	No	Comments/definitions complete enough to pass.
VER1	Reusable	Pass	Pass	No	Versioning related check passed.

Table 18: FOOPS assessment results for ERCT V5, comparing the baseline V5 run (overall score 0.9000) with the final remediated run (overall score 0.9333), reported as Pass or Failure per check.

5.8 Cross-Domain Evaluation of Interaction Modalities

Beyond international development, similar interaction challenges arise in domains that expose users to large, heterogeneous knowledge repositories. As the InterDev platform evolved to incorporate LLM-mediated exploration, it was decided to examine whether these interaction modalities are feasible beyond the specific context of international development data. While the core evaluation of InterDev focuses on usability within its target domain, understanding how similar interaction patterns behave in other complex information environments can provide additional insight into their design trade-offs. To explore this, a complementary study was conducted in the domain of digital libraries, whilst the author of this thesis was on internship in Pennsylvania State University, developing and evaluating a system that implements a comparable combination of retrieval-augmented LLM interaction and interactive graph-based navigation. The purpose of including this study in the evaluation **is not to claim generalisation**, but to assess whether the interaction modality underpinning InterDev shows signs of broader feasibility when applied to another heterogeneous data corpus. This cross-domain perspective provides a rich basis for interpreting user behaviour with LLM-assisted knowledge exploration and informs later design decisions regarding trust calibration, workload management, and explainability in InterDev. To examine whether the interaction patterns observed with InterDev reflect broader properties of LLM-assisted exploratory systems, this comparative study in a digital-library setting was conducted and the relevant findings reported here insofar as they inform RQ3.

During an internship at Pennsylvania State University, the author of this thesis designed and implemented the GutGet backend in full (Flask service, RAG pipeline, intent parsing/ontology constraints, retrieval and deduplication logic, LLM relevance gating, and graph construction). The frontend interfaces (C1–C3) was developed in collaboration with a Pennsylvania State University colleague, who co-implemented the UI components and interaction layer whilst the author of this thesis Integrated the interfaces with the backend services and authored the retrieval/LLM orchestration. In addition the research methodology and study protocol (tasks, measures, counterbalancing, and analysis plan) was designed by the author of this thesis. Data collection was conducted by a five-person study team; with the author this thesis conducting a share of the supervised laboratory sessions

and the remaining sessions were split among the other team members under the same protocol. This protocol is available in Appendix B. Ethics approval for this study was provided by the Pennsylvania State University Internal Review Board (IRB).

5.8.1 Introduction

In recent years, digital libraries and large-scale text repositories have made vast collections of human knowledge accessible, yet this accessibility introduces new cognitive and interaction challenges for end-users. A key user challenge is that such interfaces are often only optimized for keyword search and categorical browsing, which are misaligned with exploratory or curiosity-driven engagement (Carevic et al., 2018; Papenmeier et al., 2021). This challenge points to a broader distinction in information science between lookup search, characterized by specific, well-defined goals, and exploratory search, which involves more ambiguous and open-ended inquiries, emphasizing learning, sense-making, and discovery (Athukorala et al., 2016; Marchionini, 2006; Pearce et al., 2011). To ground these challenges in a concrete domain, this cross-domain research focused on large digital libraries as a testbed for studying how emerging AI-driven interfaces transform exploratory information behaviour and user cognition.

Project Gutenberg (Project Gutenberg Literary Archive Foundation, 1971) is one of the earliest and most extensive digital repositories of public domain literature, providing open access to over 75,000 texts spanning diverse genres and historical periods. Users can read books online or download them at no cost, and the collection's scale and longevity have made Project Gutenberg a widely used resource in research (Gerlach & Font-Clos, 2020; M. Jiang et al., 2022; Riddell, 2022). However, as the scale of such repositories grows, users face new challenges in interacting with large, heterogeneous text collections. Beyond discovering individual works, users often have additional ancillary requirements. They often need to interpret, compare, and synthesize information from many sources, thereby requiring iterative exploration and sense-making rather than straightforward retrieval (Boon et al., 2024). Traditional retrieval interfaces are designed for isolated search tasks, assuming that users can clearly state their goal and query a large database (Butcher et al., 2011). In practice, users refine their objectives over time, access multiple sources with varying formats, and interleave search with analysis as part of a larger interpretive process (Boon et al., 2024). However, when engaging with vast digital databases users face

fragmented information, shifting objectives, and cognitive strain as they attempt to construct a coherent understanding across sources (Boon et al., 2024; Carevic et al., 2018).

LLM-based conversational interfaces are increasingly used for exploratory search, and hybrid designs combining conversation with visualisation and clustering are emerging to support multi-document sensemaking (Mo et al., 2025; White, 2024). However, conversational search can also introduce reliability risks and cognitive overhead, placing greater emphasis on verification and confidence calibration (Gero et al., 2024). This study examines how these trade-offs play out in a digital-library setting by comparing four interface conditions during literary discovery and comprehension tasks.

Consistent with RO3, the study concentrated on two evaluation foci:

- (i) how interface modality (keyword, semantic, conversational, conversational-graph) influences task performance and strategies for discovery and synthesis; and
- (ii) how these modalities affect workload, confidence and its calibration, and perceived trust during exploratory tasks.

To explore RO3 in a setting outside international development, three LLM-mediated exploratory interfaces were implemented alongside a baseline keyword search (C0) over Project Gutenberg, a large digital-library corpus. Similar to iteration 3, a within-subjects Latin Squares condition based approach was used. The LLM-mediated variants were: **Semantic Search (C1)**, which accepts natural-language queries and retrieves books by semantic similarity rather than exact term matching; **Conversational Search (C2)**, a chat-style interface that supports interactive, turn-based refinement; and **Conversational-Graph Search (C3)**, which augments C2 with a relation-graph visualisation to support comparison and sense-making across retrieved items.

Consistent with RO3, how these modalities affect user performance and experience during representative lookup and exploratory tasks were evaluated. As in iteration 3, the study protocol measured both objective outcomes (e.g., task success and timing) and subjective experience (e.g., perceived workload, confidence, and trust calibration), using a mixed-methods design that combined quantitative metrics with qualitative observations. Results

from this study are reported only insofar as they inform the design and evaluation of InterDev’s LLM-mediated interaction approach.

5.8.2 System Design

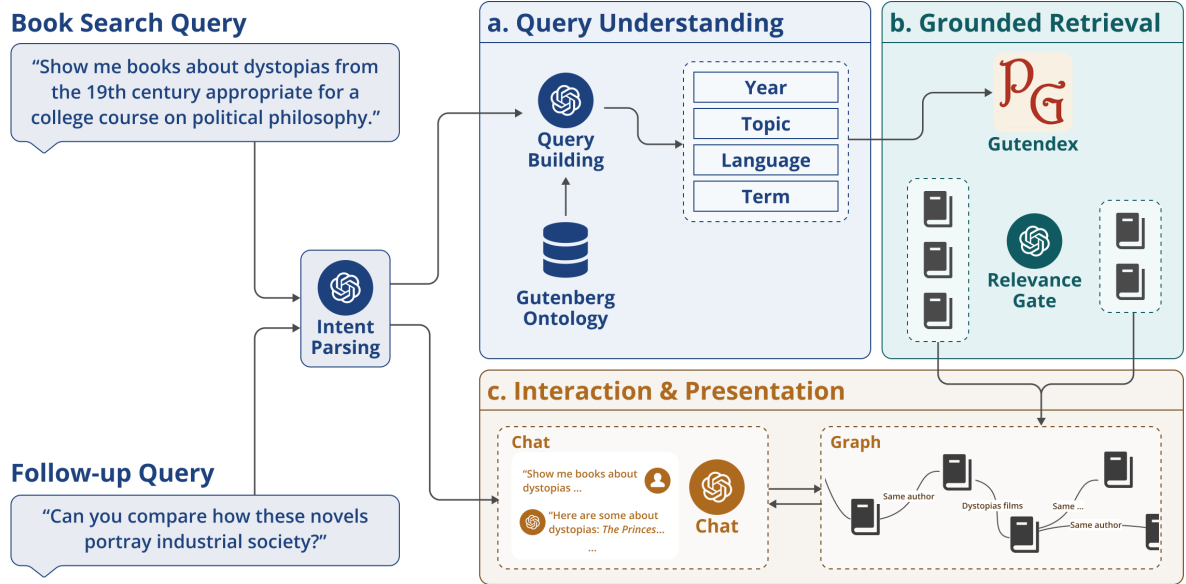


Figure 32: The GutGet backend pipeline. User prompts are parsed into structured intents constrained by a Gutenberg ontology (a), used to retrieve and filter book metadata from Gutendex with an LLM relevance gate (b), and presented through grounded explanations and an interactive graph with chat (c).

This section describes the shared backend used to support the three LLM-mediated interfaces evaluated in this cross-domain study. Beyond the original Project Gutenberg website employing keyword-based interaction (C0), three interfaces were implemented over a common retrieval-and-grounding pipeline: semantic search (C1), conversational search (C2), and conversational-graph search (C3). Using a unified backend keeps the evidence source and retrieval logic constant across C1–C3, so observed differences can be attributed primarily to interaction modality rather than to changes in underlying content access (Figure 32).

5.8.2.1 GutGet: Grounded Retrieval-Augmented Recommendations Over Project Gutenberg

GutGet is the backend engine powering conversational and graph-based book discovery. It is implemented as a Flask web service that coordinates interactions between the user interface, the Gutendex API (Gareth B Johnson, 2025), an open-access metadata service for Project Gutenberg, and OpenAI’s GPT-4o models. In keeping with linked data principles, GutGet treats each retrieved record as a graph of attributes and relationships,

constructing a lightweight knowledge graph of books, authors, and subjects at query time through Gutendex, which is itself built upon the knowledge graph expression of Project Gutenberg data, regularly published by the project as RDF (Egloff et al., 2020). GutGet employs a RAG pipeline (P. Lewis et al., 2020) that grounds every recommendation in this real metadata from Project Gutenberg.

As in iteration 3, this design mitigates a key limitation of LLMs in search contexts: hallucination, where plausible but non-existent books might otherwise appear (Ji et al., 2023). By constraining intent parsing to a fixed ontology, performing deterministic API calls, and enforcing JSON-only identifier selection, GutGet increases the probability that recommendations are real, accessible, and auditable. The pipeline produces grounded natural-language explanations alongside an interactive knowledge graph, enabling both breadth-first exploration and depth-oriented reflection.

Intent Parsing and Query Structuring

When a user submits a natural-language query (e.g., “*Show me novels by women writers in the 19th century...*”), a lightweight LLM (GPT-4o-mini) extracts a structured representation of the query aligned to books, authors, and thematic classifications for the Project Gutenberg Knowledge Graph²⁷. It specifies (i) optional temporal constraints on author birth years, (ii) up to three topical categories, (iii) the target languages, and (iv) short search terms such as exact titles or author names (Figure 32 (a)). To reduce drift, topical categories are restricted to a fixed ontology derived from Project Gutenberg’s subject and bookshelf labels, which are preloaded from a CSV file at system startup²⁸. The relevant section of this ontology are provided to GPT-4o mini as part of the prompt, with instructions to constrain the query to those relevant terms such as book categories, or else return a null result if the user’s query cannot be parsed to any existing category.

```
{
  "languages": ["en"],
  "topics": [
    "English fiction -- 19th century",
    "Women authors",
    "Social problems -- Fiction"
  ],
  "author_year_start": 1800,
  "author_year_end": 1899
}
```

²⁷ https://www.gutenberg.org/ebooks/offline_catalogs.html#xmlrdf

²⁸ <https://github.com/Xuyk021/PPGP>

}

This design provides two key benefits. First, it constrains the LLM’s interpretive flexibility to a controlled vocabulary, ensuring that user queries map consistently onto valid metadata fields. Second, by producing a machine-readable intermediate representation, the pipeline creates a transparent interface between conversational input and downstream retrieval, allowing queries to be audited, logged, or re-executed deterministically.

Candidate Retrieval and Selection.

Once an intent representation is available, the system constructs a small set of deterministic queries to the Gutendex API, each targeting different facets of the intent, such as topical categories, search terms, or author-year ranges. Each query retrieves a batch of book metadata such as titles, authors, subjects, summaries, and download links, which GutGet normalises into a node–edge structure, merging duplicates using stable identifiers. Results are then merged with duplicate suppression based on book identifiers and author–title keys to avoid redundant recommendations. To balance relevance and diversity, the backend combines two complementary sampling strategies: (i) a popularity-based ranking using download counts, which surfaces canonical works, and (ii) a topic-seeded random selection, which ensures coverage across the subjects specified in the intent. This hybrid approach prevents narrower topics from being overshadowed by globally popular titles, providing users with both familiar and exploratory options.

Before finalizing the result set, the system applies two additional safeguards. A lightweight LLM “relevance gate” filters out candidates that do not semantically match the user’s request (Figure 32 (b)). The surviving candidates are then passed through an IDs-only selection step, where the LLM is strictly instructed to return exactly n integer identifiers. If this step fails or under-produces results, the system deterministically backfills from the candidate pool. Together, these measures guarantee that every recommended item corresponds to a real, retrievable book.

Grounded Explanation and Graph Construction. As seen in Figure 32 (c), the final set of selected books is passed back to GPT-4o along with their metadata to generate a user-facing explanation. Prompts explicitly instruct the model to justify recommendations in terms of available fields, such as titles, authors, and subjects, rather than inventing content. This grounding strategy not only reduces hallucinations but also provides users with an

interpretation for why specific books were selected, strengthening confidence in the system’s recommendations.

In parallel, the retrieved books are instantiated as nodes in a force-directed graph that supports interactive exploration. Relationships between books are inferred by GPT-4o but structured to remain legible: rather than dense unstructured links, GutGet constrains relations into semantically meaningful categories (e.g., “shared author”, “common literary period”, “similar theme”), each expressed as a labelled edge (e.g., “These books share the same author” or “Both belong to Victorian fiction”). This aggregation avoids visual clutter while foregrounding the thematic or authorial connections most useful to readers. Each node is also enriched with a concise, fifteen-word summary to enable rapid scanning.

The same RAG pipeline supports follow-up interactions. When users select one or more books from the graph and issue additional queries (e.g., “How do these two novels differ in tone?”), the system retrieves and reinjects the results into GPT-4o prompts. This ensures that responses remain grounded in the selected works, enabling focused comparisons, thematic exploration, or higher-order synthesis across multiple texts. This preserves continuity across graph exploration steps, enabling iterative traversal and comparison within a semantic graph rather than isolated lists of results.

The selected metadata is injected back into a GPT-4o prompt that asks the model to generate a user-facing response grounded in the retrieved data on the user’s initial query, a variable number of results are selected and woven into a structured, paragraph-level reply that describes the books and explains why they were chosen in context. This metadata is immediately converted to graph form for relational reasoning and provenance tracking. Edges between nodes represent relationships such as shared authorship. The full prompt structure is available in Appendix C.

5.8.2.2 C0: Keyword Search

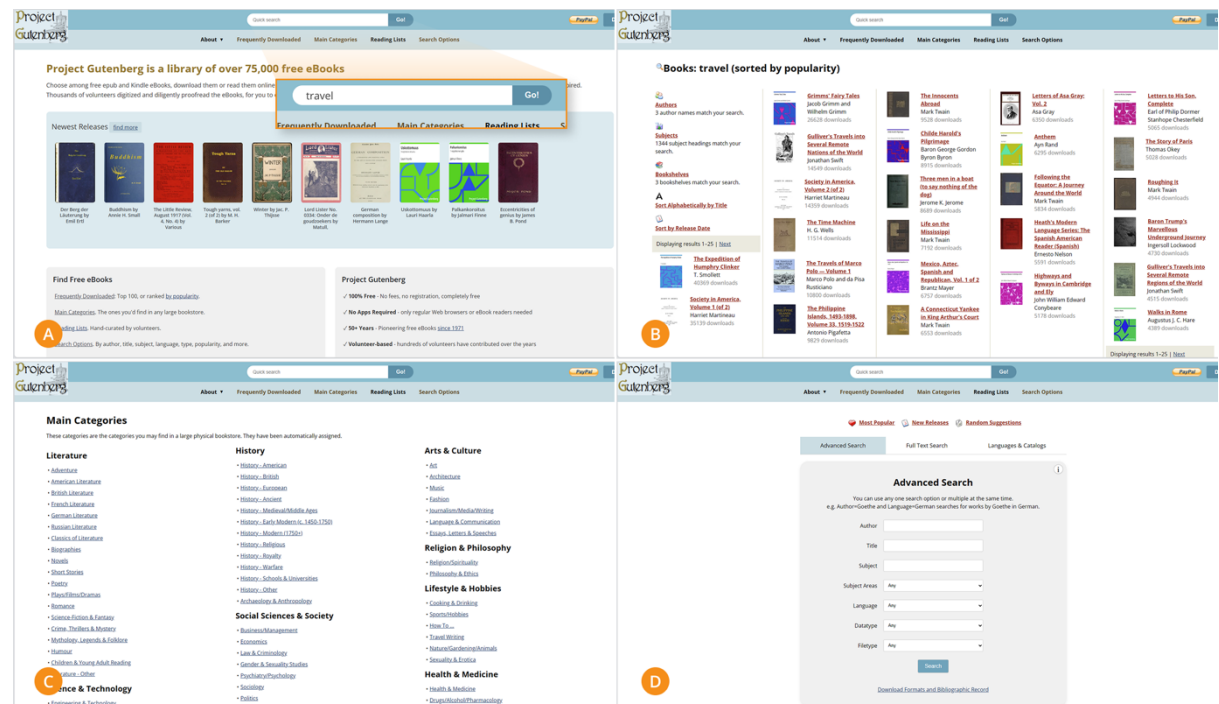


Figure 33: Overview of the original Project Gutenberg interfaces: (A) Homepage highlighting keyword search bar, (B) Example of a search results page (searching "travel", sorted by popularity), (C) Main categories listing covering literature, history, arts, and sciences, (D) Advanced search interface allowing multi-field queries (author, title, subject, language, etc.).

The Project Gutenberg interface (Figure 33) primarily relies on keyword-based search as its main retrieval mechanism. Users can input fragments of a title, an author's name, or other textual cues to retrieve relevant books from the digital collection (Figure 33B), making it a helpful form of interaction on the platform (Bates, 2002; White, 2016).

In addition to keyword search, the website offers category browsing and structured search options (Figure 33C–D). Category browsing organizes books by subject or genre, while structured search enables field-specific queries using predefined metadata such as author, title, or subject. Selecting any result opens its detailed book page (Figure 34).

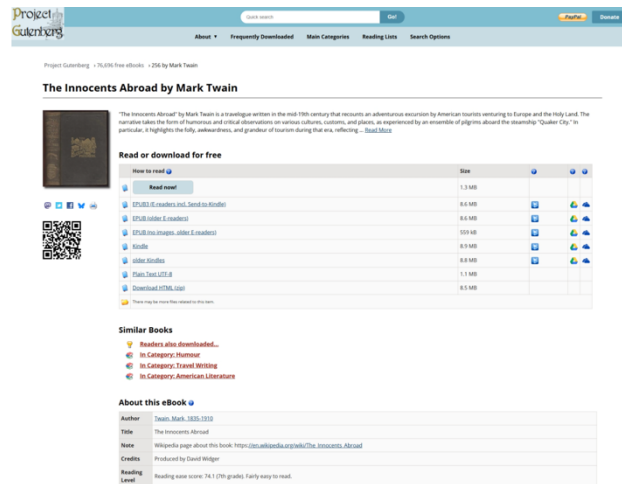


Figure 34: Overview of the original Project Gutenberg website interfaces of the book detail: introduction, download links, categories and other detailed data.

5.8.2.3 C1: Semantic Search

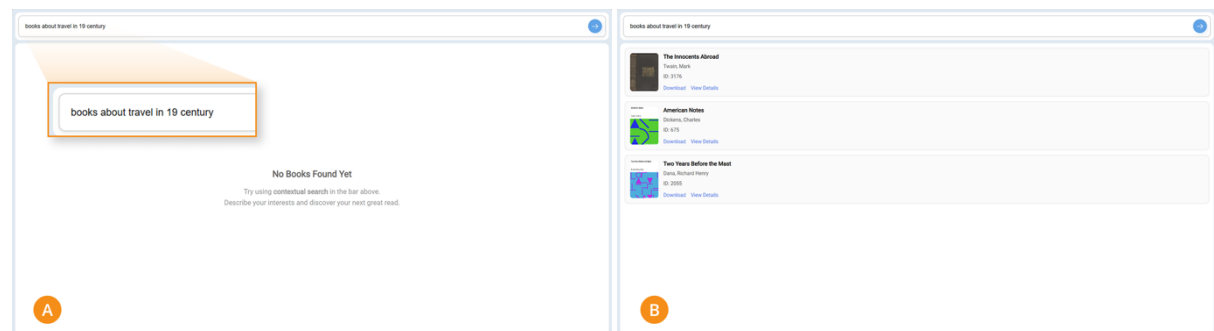


Figure 35: Overview of our Semantic Search and an example query results for “books about travel in 19th century”: (A) no results returned with keyword search before searching, (B) relevant results retrieved with contextual search.

Semantic search has become a common information retrieval method in recent years across different domains, such as code search (X. Li et al., 2021) and data repositories search (Setlur et al., 2023). In this interface, users can enter queries in natural language form to find books, and the system retrieves semantically related results rather than relying on exact word matches, while the overall query process and result presentation remain similar to keyword-based search (Linckels & Meinel, 2011). Examples include “books about fishing in the eighteenth century”, “novels set in Victorian London”, etc. The system processes this text input and returns a list of books that have strong semantic relevance to the query, as shown in Figure 35.

5.8.2.4 C2: Chat Search

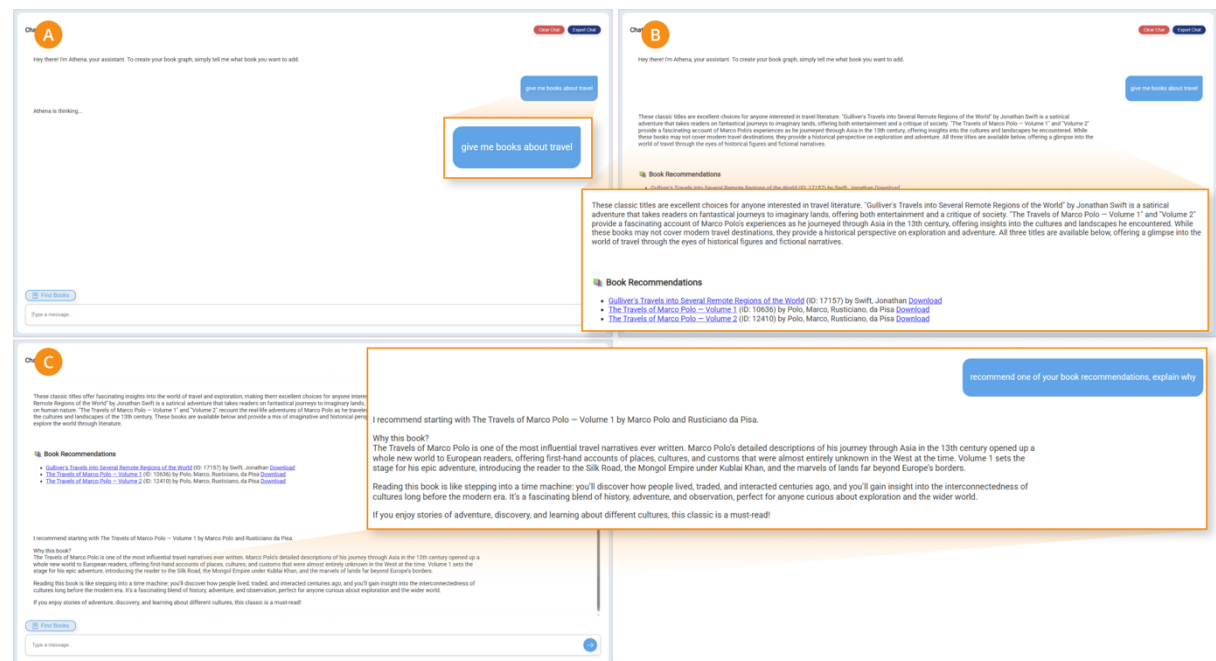


Figure 36: Overview of the Conversational Search interface: (A) User initiates a query ("give me books about travel"), (B) System responds with book recommendations and hyperlinks, (C) User follows up with another request ("recommend one of your book recommendations, explain why"), (D) System provides a detailed explanation supporting the recommendation.

In this interface, users can directly type their questions or information needs into the chatbot (Figure 36). The system responds with relevant book results accompanied by brief explanations. For example, when a user asks "Give me books about travel", the system returns a set of recommended titles. Users can then continue the dialogue with follow-up requests, such as "Recommend one of your book recommendations, explain why", to which the system provides a detailed rationale highlighting the book's historical and cultural significance. This illustrates how conversational interaction enables iterative refinement and supports exploratory engagement within the conversational threads.

5.8.2.5 C3: Conversation-Chat Search

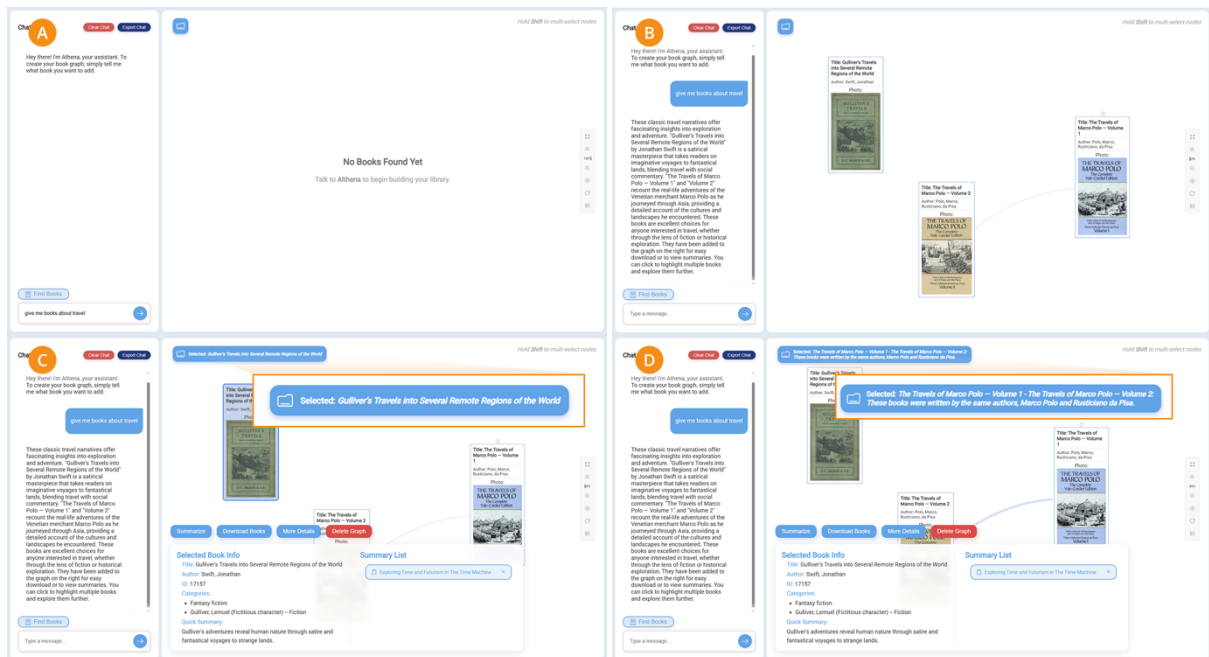


Figure 37: Overview of our Conversational-Graph Search Interface (C3) with an example query “books about travel”: (A) initial query with no results yet, (B) contextual results retrieved and visualized in a relation graph, (C) user selection of one book and showing the metadata of the book, (D) system highlighting relationships between selected books.

C3 integrates a real-time graph visualization into the conversational interface. Each recommended book is represented as a node, and edges represent relationships between books based on inferred semantic similarity, shared authorship, literary period, or related thematic elements. Users can also click nodes to expand nearby recommendations, browse summaries, and download books into personal collections. This synchronized conversational and visual graph approach was designed to allow users to fluidly move between conversational interaction and graph exploration, combining breadth of discovery with depth of reflection.

5.8.3 User Study

To evaluate the interface modalities a within-subjects laboratory study was conducted to examine how different forms of system support shaped book discovery, interpretation, and synthesis in Project Gutenberg. The design was intended to capture both objective task outcomes and subjective user experience, while minimizing individual differences across participants. Participants were compensated \$20 upon completion of the study.

Ethical approval was granted by the Internal Review Board (IRB) of Pennsylvania State University.

5.8.3.1 Participants

Participants were recruited through Pennsylvania State university mailing lists, campus posters, and local networks. Recruitment was stratified by age, gender, and education to ensure a balanced participant pool. Twenty-eight individuals completed the study ($n = 28$). The sample included 19 females, 7 males, 1 non-binary participant, and one who did not report gender. Age was broadly distributed, with participants spread across 18–24 (6), 25–34 (6), 35–44 (6), 45–54 (5), and 55–64 (5) years. Education levels ranged from Associate (3) and Bachelor’s (12) to Master’s (8) and Doctoral (5) degrees. Seventeen participants reported English as their first language. Reading frequency was generally high: 14 described themselves as “very frequent” readers and another 14 as “somewhat frequent.” Familiarity with LLMs (1–5 scale) was moderate to high (median = 3, range = 1–5). Thirteen participants mentioned prior awareness of digital libraries such as ACM, IEEE, Web of Science, or Libby.

A within-subjects design was employed so that each participant experienced all four user interfaces. The order of interface presentation was counterbalanced using a 4×4 Williams Latin square to mitigate sequence and learning effects (MacKenzie, 2024). Participants gave informed consent and received a \$20 gift card upon completion of the study as compensation.

5.8.3.2 Conditions

Four user interfaces that provided progressively greater levels of system intelligence and exploratory support, were compared:

- **C0: Project Gutenberg (baseline).** The unmodified Project Gutenberg website supporting keyword, category, and structured search. Results are presented as a ranked list of matches with standard metadata (title, author, year) and links to the full text. No model-based re-ranking, rationales, or conversational affordances are provided.

- **C1: Semantic search.** A natural-language retrieval interface that uses the same retrieval and grounding backend as C2 and C3 as described in Section 5.8.2 but exposes results in a non-conversational list view. Rankings are produced by the retrieval and LLM re-ranking pipeline with zero temperature and a fixed seed; no conversational history or turn-based context is preserved between queries. Participants may submit additional queries, each treated as a fresh request.
- **C2: Conversational search.** A chat-style interface that accepts natural-language questions and returns recommended works per response. Each suggestion is accompanied by a short explanatory rationale (1–3 sentences). The conversational interface preserves turn-wise history to allow follow-up questions and targeted clarification. The search layer and backend is identical to C1 as described in Section 5.8.2; the manipulation is therefore the conversational interaction model and the addition of explicit rationales.
- **C3: Conversation-Graph search).** Extends C2 with an interactive relation graph visualization. Edges indicate pairwise relations derived from the same grounding signals used by the retrieval pipeline described in Section 5.8.2. Users can click nodes to reveal the chat context and evidence that supported the recommendation, expand neighbours up to depth $D = 2$, and traverse the graph.

Each interface was introduced through a short tutorial task that familiarized participants with its capabilities. Tutorial outputs were not analysed. The full task list is available on the thesis GitHub under the cross-domain study folder²⁹.

5.8.3.3 Study Procedure

Eight evaluation tasks covered three key themes in exploratory search: (i) lookup and constraint-based filtering (Hearst, 2009; Marchionini, 2006), (ii) interpretation and verification (M. J. Metzger, 2007; Pirolli & Card, 2005), and (iii) synthesis and comparison (Athukorala et al., 2016). All tasks were interface-agnostic and could therefore be solved with any interface. For each participant, within each interface block, two distinct tasks were randomly sampled without replacement from the common task pool (yielding eight tasks per participant, two per interface). Sampling was stratified by theme to avoid theme-level

²⁹ <https://github.com/mmurtaghw/InterDevResearch>

imbalance within participants. Across participants, each task therefore appeared under multiple interfaces, enabling estimation of overall interface main effects and task-family-specific contrasts.

Task	Family	Goal and constraints
T1	Lookup/Filter	Find the book <i>Frankenstein</i> by Mary Shelley and retrieve the exact opening sentence.
T2	Lookup/Filter	Locate a Jules Verne novel available in plain-text format that satisfies the task’s download constraints.
T3	Interpret/Verify	Extract author, dedication, and intended audience details from Machiavelli’s <i>The Prince</i> , citing the required metadata.
T4	Interpret/Verify	Select a short science-fiction book that satisfies provided constraints and justify the choice.
T5	Synthesis/Comparison	Recommend a pair of political-philosophy works that meet complementary criteria and explain the pairing.
T6	Synthesis/Comparison	Assemble a mini syllabus on utopia/dystopia themes with rationale spanning the requested perspectives.
T7	Interpret/Verify	Judge the truth of a Mary Shelley claim and support the verdict with evidence from her works.
T8	Synthesis/Comparison	Compare industrial-anxiety narratives across two titles, highlighting similarities and differences.

Table 19: Evaluation tasks and their target behaviours. Each participant completed two distinct tasks per interface block.

Sessions lasted approximately 90 minutes and were supervised by a member of the research team. All materials (tasks and questionnaires) were delivered via Qualtrics survey. After providing consent, participants first completed a short demographic questionnaire, after which they proceeded through four interface blocks whose order was counterbalanced using a 4×4 Williams Latin square.

5.8.3.4 Data Processing

All task outputs were exported from Qualtrics and processed through a scripted pipeline. Because factual tasks required exact matches, a rubric containing the expected strings (e.g., the first sentence of *Frankenstein*, the catalog of acceptable Jules Verne titles) was constructed. Participant answers were normalized through case folding and whitespace collapse before being matched against the key; each field was marked as satisfying or violating the task constraints. The same workflow produced descriptive feedback that helped us diagnose common errors. This process was reviewed manually by a member of the research team.

Open-ended synthesis tasks demanded a more nuanced assessment. Therefore three-criterion rubric was designed, capturing (i) whether the response met the stated constraints, (ii) the correctness of supporting snippets, and (iii) the depth and coherence of the rationale or thematic integration. Two trained raters scored all submissions according to the rubric and were aggregated.

5.8.3.5 Measures

To evaluate how each interface influenced user performance and perceptions, objective task outcomes with subjective measures of confidence, workload, and trust were combined. Some of these measures were shared with iteration 3, those which are new or have different configurations are outlined below:

Task performance. Similar to iteration 3, each task was scored on its stated constraints: lookup and filter outputs were marked as 1 when all explicit requirements were satisfied (e.g., exact sentence, valid author/title, correct verdict) and 0 otherwise, while synthesis and interpret and verify-style prompts used the rubric dimensions (constraint fit, snippet/evidence correctness, rationale depth) to generate summed ordinal scores. Two trained raters annotated the synthesis and interpreted and verified items via our rubric. Finally, scores were averaged for analysis. Agreement was strong, with a Cohen's $\kappa = 0.88$.

Qualitative feedback. Each participant completed a short post-condition reflection and a final post-study interview, providing insight into strategies, breakdowns, and perceptions of each interface.

5.8.3.6 Analysis

All quantitative outcomes were analysed using mixed-effect regression models, treating interface condition as a fixed effect and accounting for repeated measures by subject. Tasks entered all generalized linear and mixed-effects models as fixed effects. Task completion time was analysed using a log-linear mixed model that included participant random intercepts; log-time models were likewise tested with task random intercepts, which produced negligible changes in model estimates. Synthesis/comparison and verify/interpret rubric scores were examined with a cumulative-link mixed-effects model using a logit link

and participant random intercepts. Self-reported confidence (0–100, rescaled 0–1) was modelled using a linear mixed-effects model on logit-transformed scores, with fixed effects of interface condition and task type and participant random effects. Confidence, calibration, and Brier score analyses also confirmed that results were stable when including additional task random effects. This modelling approach follows standard practice in quantitative HCI for repeated-measures user studies, where mixed-effects models are recommended to account for within-participant dependence and other grouping structures that are common in interaction data (Robertson & Kaptein, 2016). These can create non-independent observations that violate the assumptions of simple ANOVA or ordinary regression. Participant was therefore modelled as a random effect to account for within-subject dependence, while interface condition and task were included as fixed effects to estimate condition differences while controlling for task difficulty. Model families were selected to match outcome structure: log-linear models for positively skewed completion times, cumulative-link models for ordinal rubric scores, and linear mixed models for approximately continuous confidence measures after logit transformation.

Interview data were coded independently by two authors using thematic analysis (Braun & Clarke, 2006), in line with similar qualitative analyses in the domain (Brachman et al., 2025; X. Wang et al., 2025), as described in section 5.3.3.2. The team collaboratively developed a shared codebook, resolved discrepancies, and refined themes through discussion.

5.8.4 Results

This section presents quantitative and qualitative evidence on how the varying modalities shapes users' search behavior, task performance, and subjective experience. First task outcomes, workload, efficiency, and confidence patterns across the four conditions are reported, followed by participant reflections to provide insight into the mechanisms behind these effects.

5.8.4.1 Quantitative Results and Analysis

Task Performance. Higher scores indicate better performance unless otherwise noted. Synthesis tasks benefited from the conversational interfaces (Figure 38, left panel). Relative to the Project Gutenberg baseline (C0), the chat-only condition (C2) yielded a positive

effect ($\beta = 0.225$, $p < 0.001$) and the conversational-graph condition (C3) produced a comparable gain ($\beta = 0.217$, $p < 0.001$). The semantic list (C1) delivered a modest but reliable improvement over the baseline ($\beta = 0.106$, $p = 0.043$). C2 tended to outperform C1 ($\beta = 0.118$, $p = 0.085$), while C2 and C3 performed similarly.

Interpret/verify tasks showed a different pattern (Figure 38, center). Mean scores ranged from 3.5–4.9 on the 0–7 rubric; the conversational-graph interface produced the highest average (4.9) and significantly exceeded both the semantic list (C1, $\beta = 0.490$, $p = .031$) and chat-only (C2, $\beta = 0.404$, $p = .048$) conditions, although differences versus the baseline did not reach significance.

Lookup/filter accuracy was near ceiling for every interface (Figure 38, right). Both the baseline and conversational-graph conditions achieved 100% accuracy, while the semantic list and chat-only interfaces remained high (90% and 94%, respectively), indicating that all designs supported the basic lookup tasks effectively.

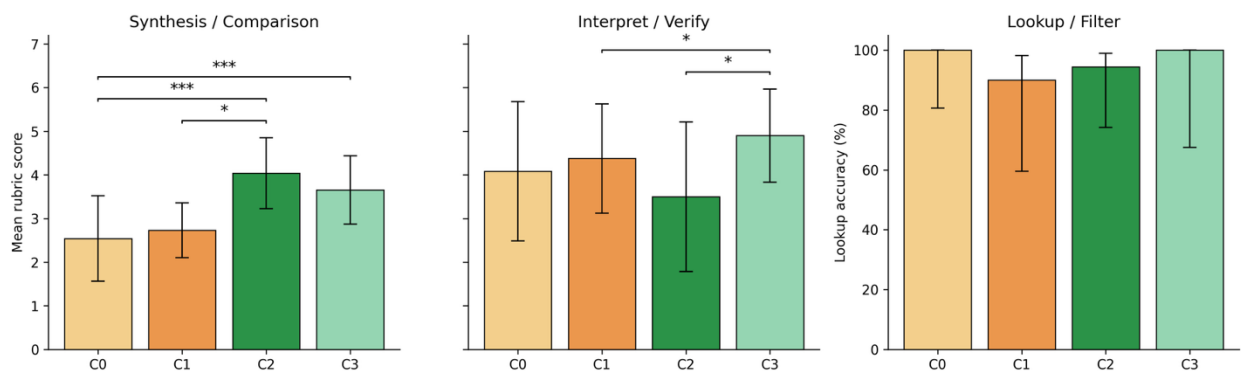


Figure 38: Mean outcomes by interface condition. The first two panels show rubric scores for synthesis/comparison and interpret/verify tasks (0 – 7), the right panel reports lookup/filter accuracy. Error bars represent standard errors (* $p < .05$, ** $p < .01$, *** $p < .001$)

Workload (NASA-TLX). Figure 39 shows that interface design also shaped perceived workload. Relative to the baseline (C0), the chat-only interface (C2) produced the largest reductions across subscales: mental demand ($\beta = -2.32$, $p < 0.001$), temporal demand ($\beta = -1.48$, $p = 0.001$), effort ($\beta = -2.28$, $p = 0.001$), and frustration ($\beta = -2.13$, $p = 0.004$) all dropped while perceived performance increased ($\beta = +1.96$, $p = 0.005$). The conversational-graph interface (C3) preserved the performance boost ($\beta = +1.40$, $p = 0.041$) but did not significantly reduce the other subscales. The semantic list (C1) delivered a modest temporal-demand reduction ($\beta = -1.13$, $p = 0.010$). As expected, physical demand did not differ across conditions.

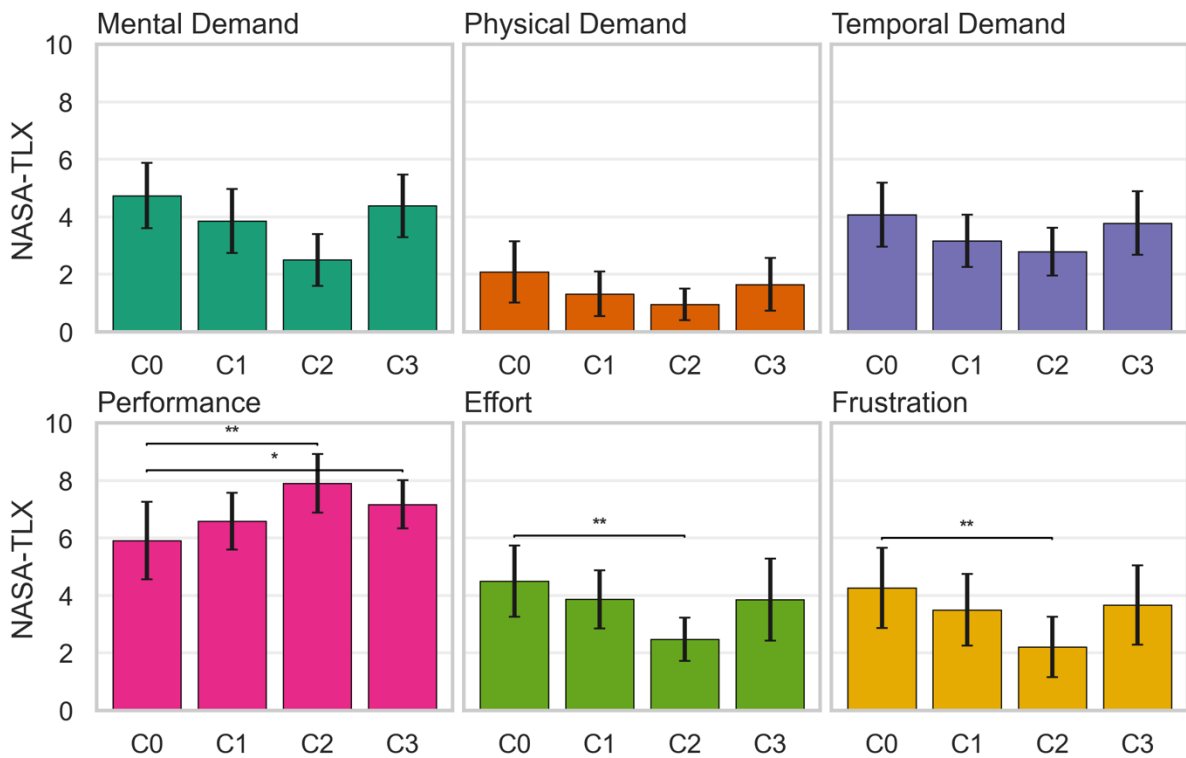


Figure 39: Mean rubric synthesis task scores (0–7). Error bars represent standard errors (* $p < .05$, ** $p < .01$, *** $p < .001$)

Completion Time. Interface condition had selective effects on efficiency (Figure 40). Aggregated across all tasks, the chat-only interface (C2) was significantly faster than the baseline (C0), reducing observed completion time by approximately 109s ($p = 0.0499$). The semantic list (C1) and conversational-graph (C3) were not reliably different from C0, and C2 also trended faster than C3 ($p = 0.066$). Breaking results down by task family with mixed-effects models on log-time, complementary patterns are observed. For Interpret/Verify tasks, both C1 ($\beta = -0.56$, $p = 0.007$) and C2 ($\beta = -0.62$, $p = 0.024$) were faster than C0, while C3 showed no reliable difference. For Lookup/Filter, C1 increased time relative to C0 ($\beta = +0.58$, $p < 0.001$), whereas C2 and C3 did not differ from baseline. For Synthesis/Comparison, no condition differences reached significance (all $p > 0.05$). Overall, chat support accelerates evidence-gathering tasks, the semantic list can slow simple lookups, likely due to additional query refinements, and the graph layer does not

deliver efficiency gains; synthesis remains time-intensive regardless of interface.

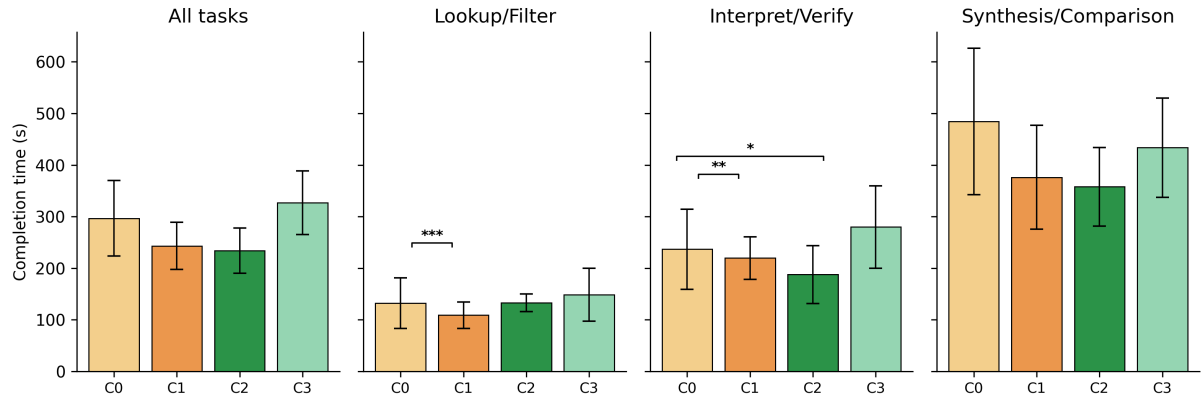


Figure 40: Mean completion time across task types. Error bars represent standard errors (* $p < .05$, ** $p < .01$, *** $p < .001$)

Figure 41 shows that the richer interfaces substantially raised participants' confidence, while calibration remained steady. A linear mixed-effects model on the logit of reported confidence, with participant random intercepts and task fixed effects, found that all conditions beyond the Project Gutenberg baseline (C0) produced higher confidence: Semantic (C1) $\beta = +2.23$, $p = 0.017$; Chat (C2) $\beta = +2.76$, $p = 0.0027$; conversational-graph (C3) $\beta = +3.45$, $p < 0.001$. Task-type fits reveal where the uplift occurs. In Interpret and Verify tasks, C3 significantly exceeded C0 ($\beta = +4.22$, $p = 0.011$) and C1 trended upward ($p = 0.052$); in Synthesis and Comparison tasks, C2 approached significance ($\beta = +3.12$, $p = 0.052$). Lookup and Filter tasks were near ceiling for every interface, yielding no reliable differences.

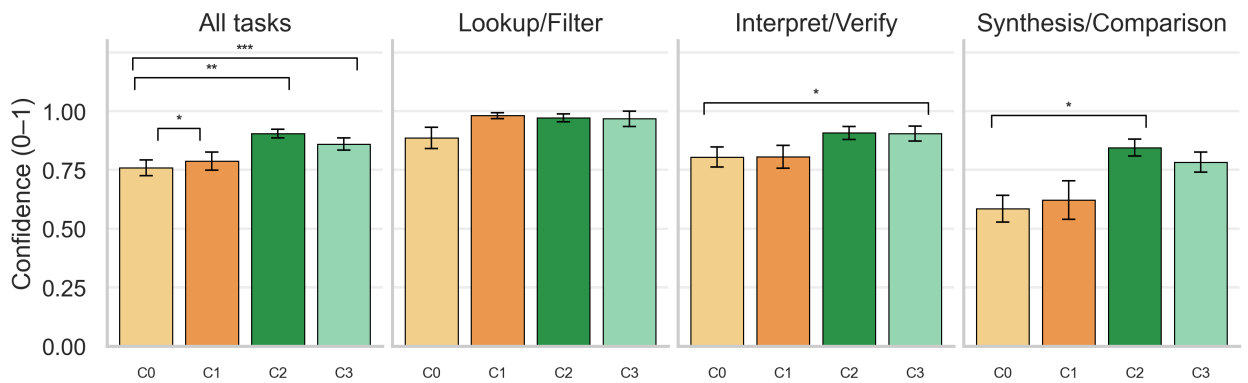


Figure 41: Self-reported confidence scores by task type. Error bars represent standard errors (* $p < .05$, ** $p < .01$, *** $p < .001$)

Confidence calibration is complementary to this. Calibration is quantified as the mean squared error between reported confidence and correctness (Brier; lower indicates closer

calibration). Aggregated across all tasks, the participant-level mixed-effects model found no reliable differences from the keyword search baseline (Figure 42): C1 ($\Delta = +0.01$, $p = 0.92$), C2 ($\Delta \approx 0.00$, $p = 0.99$), and C3 ($\Delta = -0.05$, $p = 0.48$). Task-type models tell the same story. For Lookup/Filter, Brier scores were uniformly small (< 0.10) with no significant contrasts. Interpret/Verify elevated scores for all interfaces (C0 ≈ 0.47 , C3 ≈ 0.21), but even the largest contrast (C3 vs. C0) narrowly missed significance ($p = 0.084$). Synthesis/Comparison sat in the mid-range (≈ 0.18 – 0.31) with overlapping intervals across conditions (all $p > 0.20$). Overall, the results suggest confidence–accuracy calibration did not measurably change across conditions. The results suggest the chat-based conditions make people feel more confident without undermining alignment between their probability judgments and actual correctness.

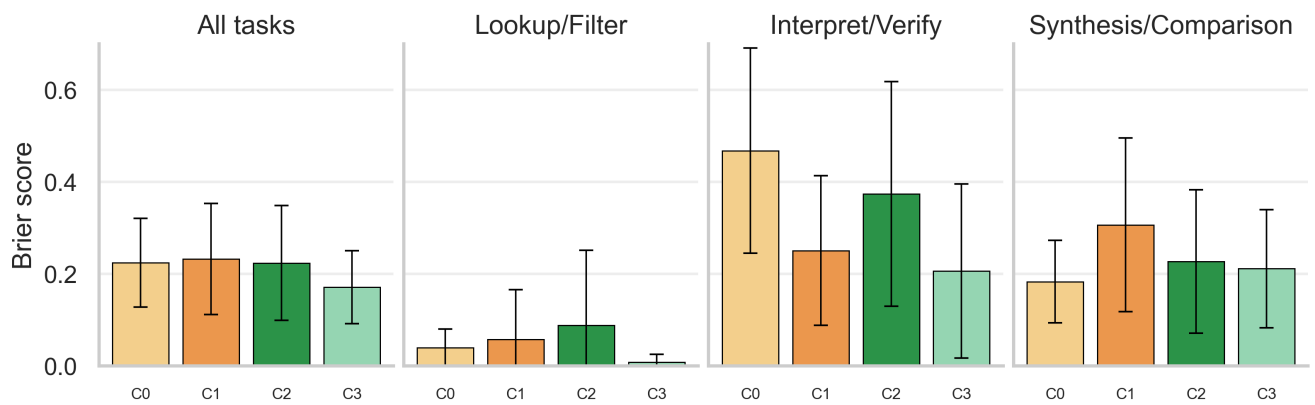


Figure 42: Brier score by task type. Error bars represent standard errors (* $p < .05$, ** $p < .01$, *** $p < .001$).

5.8.4.2 Qualitative Results

Interviews and open-ended responses revealed six themes that explain how participants viewed each interface and why the quantitative patterns emerged. This analysis also served to further clarify and understand the design implications of the experiment. As summarised in Table 20 these themes span query formulation, trust and verification, comparison support, output depth, stopping behaviour, and perceptions of system expertise.

Theme	Findings	References	Scope
Conversational momentum	Follow-up carryover and iterative reformulation reduced articulation effort and perceived workload during exploratory search.	38	Query formulation and workload
Verification thresholds	Users' decisions to verify depended on prior knowledge, effort required, and trust cues such as visible source provenance.	136	Trust, provenance, and checking
Comparison friction	Visual aids supported synthesis and comparison across items, but visual density and extra interaction steps increased cognitive and interaction overhead.	91	Comparison support and visual overhead
Minimal output	Concise outputs improved scanability and simplicity, but often lacked rationale, provenance, and coverage, creating hidden verification costs and limiting synthesis.	86	Scanability versus depth
Sufficiency	Interfaces shaped when users felt they had "enough" information; these stopping decisions were driven more by perceived completeness than by correctness, sometimes causing premature closure.	14	Stopping cues and perceived completeness
Expert proxy	Some participants treated the system as an expert partner, offloading cognitive work and gaining confidence even when objective accuracy did not necessarily improve.	15	Cognitive delegation and perceived expertise

Table 20: Summary of qualitative themes identified from interviews and open-ended responses, including their key findings, reference counts, and analytical scope.

Conversational momentum. Conversational momentum refers to the user's ability to reformulate queries and lower articulation cost to create gains in workload. Participants generally preferred being able to clarify with reference to previous queries without having

to restart their exploratory search. Some participants noted that this allowed them to iteratively improve their queries to get the results they needed, e.g. *“I liked it because I felt like I could ask follow-up questions... But that made me ask better questions when it started”* (P1). Others referred to less mental workload in terms of having to articulate the query: *“So I’ve just felt like I could at least, I didn’t have to put so much effort into thinking about what I was going to search so I could just start searching and getting right into it.”* (P4).

Verification thresholds. Verification thresholds refer to the situational rules or cues that prompt a user to verify their responses and when to verify system-provided claims. Many users deferred verification checks or skipped such checks when they felt their own background knowledge was sufficient. Interestingly, verification thresholds seemed to shift depending on interface context—when books were directly surfaced from Project Gutenberg by a chat-based interface some users reported feeling more trustworthy e.g. *“There was a link to Project Gutenberg, so I felt it was connected to a more trustworthy source.”* (P2) whereas other users noted scepticism without: *“I just, in general, don’t really trust AI-based models as much as I do, me finding those sources on my own.”* (P6).

Comparison Friction. Comparison friction refers to the visual trade-offs reported by users across conditions. Visual artefacts such as graphs, summaries, relation edges and book covers enabled externalised comparison and cross-item synthesis, but their visual density, layout and additional manipulation steps introduced interaction overhead that raised effort. Users reported utility in synthesis and compare-and-contrast tasks e.g. *“chat and graph one, because I can have like one click, it’s summary... then I can compare and contrast those”* (P17), but users also mentioned issues of cognitive overload: *“the graph one was too much, and I didn’t like that the actual chat window was way smaller.”* (P20).

Minimal output. Minimal output refers to concise, surface-level outputs such as titles, short blurbs or very brief chat replies that are easy to scan but omit supporting rationale, provenance or coverage for confident verification or synthesis decisions. Interestingly, a tension can be identified here with the comparison friction theme; for example, users were mindful of simplicity of the semantic search interface and that the results were comparable to other LLM-based interfaces: *“Probably the first one. [Semantic Search]... It just seemed*

the simplest... here's your question. Here's your answer..." (P24), but did not feel they had the same immediate depth of knowledge, leading to hidden verification costs: *"The one where I felt least confident was definitely the semantic one... If I didn't get it right I knew there wasn't really much I could do to confirm anything else."* (P20). These minimal outputs can be helpful for lookup and simple verification tasks, but without a clearer route to further explore the underlying literature they were less helpful in synthesis tasks.

Sufficiency. Participants highlighted the role of interfaces in signalling when they had gathered enough information to complete a task, make a decision, or stop searching. In this context, sufficiency reflects both cognitive and interface-level cues that indicate when further effort no longer seems necessary or productive. Participants described feeling clear sufficiency signals in the chat-based paradigms, where coherent and self-contained responses gave a sense of closure: *"The chat one, I felt like I got all the information I needed and I could confirm that it kind of gave me it and then I could confirm it pretty easily and quickly myself"* (P20). By contrast, sufficiency was less apparent in the semantic search condition, where participants remained uncertain whether results were exhaustive or complete: *"I would say the last one [Semantic Search] only because there was a lot there... so I wasn't confident that that information was pulling everything"* (P24). Some participants described this as a feeling of being "stuck between" searches, unsure when to stop refining their query. In the conversational-graph interface, the visual layout itself sometimes acted as a stopping cue. When all visible nodes appeared relevant, participants inferred that they had seen the whole conceptual space, even if verification was incomplete. Across conditions, sufficiency judgments appeared to arise from perceived completeness rather than objective correctness. Chat interfaces produced earlier stopping and lower workload but occasionally led to premature closure, reflecting higher subjective confidence without corresponding gains in calibration. Sufficiency therefore interacts closely with the preceding theme of Verification Thresholds: while thresholds govern whether users choose to check information, sufficiency governs when they decide to stop.

Expert proxy. Expert proxy refers to instances where participants treat the system as an expert partner and defer elements of cognitive work to it. This may result in situations where participants felt guidance and subjective confidence even if their task accuracy does not meaningfully change. Participants mentioned viewing the model as an expert that had

knowledge beyond their own: *“If you have a model that you can ask questions to, it makes you more confident to your answer because it’s not just you based upon what you can find about the books... I guess it feels like you have someone else there who’s an expert”* (P19). Participants reported offloading cognitive work to interfaces that featured a chat function, e.g. *“I could just cut and paste the question and put it into the chat for the answer instead of trying to solve the puzzle with pieces myself”* (P18), though some participants were wary of this anthropomorphism: *“It feels like it knows what it’s talking about, even though I know that we shouldn’t always trust it”* (P22).

5.8.5 Discussion

The cross-domain study provides further context for interpreting the results of Iteration 3. Although the application domains differ, both settings involve interaction with large, heterogeneous corpora and require users to perform higher-order activities such as comparison, verification, and synthesis. This supports interpreting the Iteration 3 findings not as artefacts of the ERCT corpus or task set, but as reflecting more general properties of LLM-mediated exploratory systems.

Across both studies, two complementary patterns emerge. First, conversational interaction reliably reduces interactional effort but does not remove the epistemic work of synthesis and judgment. Second, confidence and its calibration are shaped by interface cues in ways that are not straightforwardly aligned with correctness. The following sections discuss these two patterns and their implications for InterDev.

5.8.5.1 Chat as interactional offloading, with synthesis as the remaining bottleneck

Across both studies, conversational interaction consistently reduced perceived workload and interactional friction. In Iteration 3, both chat conditions lowered NASA–TLX scores relative to the baseline, particularly for mental demand, effort, and frustration, yet neither task performance nor completion time improved once task difficulty was controlled. The cross-domain study shows a similar trend. Chat-based interfaces reduced workload and, for some task families, improved efficiency, but synthesis tasks remained time-intensive and performance gains were selective.

Taken together, these results suggest that conversational interfaces primarily function as a mechanism for reducing the cost of navigation, query formulation, and cross-referencing. They do not, by themselves, reduce the epistemic effort required for interpretation and integration (L. Wang et al., 2025), and prior work shows that fluent or authoritative conversational presentations can increase perceived credibility and confidence without improving underlying judgment quality (L. Metzger et al., 2024; Xu et al., 2025). This interpretation is reinforced by the qualitative findings in both studies. In Iteration 3, participants reported that chat made comparison across trials easier, but that outputs often enumerated findings rather than integrating them into a coherent synthesis. Similarly, in the cross-domain study, conversational interaction promoted exploratory momentum and easier cross-item inspection, while deeper integration still depended on user judgment and additional effort.

This convergence helps explain why reduced workload in Iteration 3 did not translate into improved performance (Yun & Bickmore, 2025). Once access and comparison become easier, the primary bottleneck shifts to integration, reconciliation of evidence, and construction of higher-level interpretations. For InterDev, this implies that further performance gains are unlikely to come from additional retrieval or navigation support alone, but instead from more explicit scaffolding for synthesis and reasoning over curated evidence.

5.8.5.2 Balancing Exploratory Power and Cognitive Load in Hybrid Visual–Conversational Interfaces

The two studies diverge in a way that is particularly informative for interpreting Iteration 3’s results on confidence and calibration. In the InterDev study, adding citations in the chat-with-sources condition reduced self-reported confidence and worsened confidence calibration relative to baseline, while trust in automation remained stable. Qualitative data indicated that citations were treated as cues for scrutiny rather than as assurances of correctness, prompting greater caution without necessarily improving judgment quality. In contrast, the cross-domain study found that conversational interfaces increased self-reported confidence relative to the baseline, while calibration remained broadly unchanged.

Taken together, these results indicate that confidence is highly sensitive to interface cues and interaction style. Conversational fluency can create a sense of sufficiency and closure

that raises confidence, whereas visible provenance can foreground uncertainty and reduce confidence. Neither effect, however, guarantees better alignment between confidence and correctness. This sharpens the interpretation of Iteration 3. Provenance visibility should not be expected, on its own, to improve trust calibration. Citations can change users' subjective stance toward their answers without providing efficient means to resolve uncertainty, particularly when verification remains effortful or when references operate at a coarse granularity, such as the trial level rather than the level of specific claims or findings.

This comparison also reinforces the need to treat synthesis and verification as design targets. In both studies, richer structural or visual scaffolds supported comparison but introduced additional interaction costs, sometimes offsetting the workload benefits of chat (Avula et al., 2022; P. Jiang et al., 2023). For InterDev, this suggests that provenance and structural aids should be deployed selectively and in a task-aligned manner, and that calibration should be addressed through claim-level, actionable support rather than through transparency alone deployed selectively and in a task-aligned manner (Avula et al., 2022; Hoeber & Shukla, 2022). This perspective directly motivates the post-Iteration 3 move toward a finding-level evidence model, in which the unit of citation and interaction more closely matches the unit of reasoning and verification.

5.8.5.3 Limitations

While the findings offer empirical evidence for how conversational and hybrid interfaces influence confidence, workload, and trust in exploratory search, several limitations constrain generalization. First, the study was conducted in a controlled laboratory setting using a fixed task protocol and a single digital library corpus (Project Gutenberg). Although this environment allowed for rigorous within-subject comparison, it may not fully capture the variability of open-world information seeking, where goals are fluid, interruptions occur, and users interleave multiple tools and devices. Future field studies could evaluate how conversational and graph-based exploration unfold in more natural settings, such as academic literature review or cultural heritage discovery.

Second, the retrieval-augmented pipeline was implemented with a specific LLM configuration and a curated ontology. As model capabilities, alignment techniques, and retrieval strategies evolve, both the quality and transparency of explanations may shift.

Replicating the study with alternative models and retrieval frameworks would help isolate which effects arise from interface design versus model behaviour.

Third, although the mixed-methods approach provided insights, the sample size ($n = 28$) and demographic composition limit generalization across populations. Participants were relatively literate in digital search tools and moderately familiar with LLMs, which may have moderated both confidence and workload perceptions. Broader participant sampling, including novice and domain-expert groups, could clarify how prior expertise and AI awareness moderate the benefits of conversational scaffolding (Cheng et al., 2025). Additionally, while trust surfaced regularly in the qualitative results of our study, quantitative measures of trust such as Jian et al. (Jian et al., 2000) or Malle and Ullman (Malle & Ullman, 2021) were out of the scope of the experiment as they would have added considerable length and cognitive load to the study protocol. A follow-up, larger study using these measures might prove informative about patterns raised in the qualitative portion of our research.

Finally, the study examined short, session-bounded interactions. Longitudinal work is needed to understand how repeated exposure affects confidence calibration, trust repair, and reliance over time. Extending the system to support persistent memory or cross-session learning could enable such investigation while raising new questions about transparency, provenance, and ethical personalization in exploratory AI systems.

5.8.6 Cross Domain Evaluation: Conclusion

This study examined how different interface modalities for book discovery influence user confidence, workload, and trust in LLM-powered exploratory search. By comparing keyword, semantic, conversational, and conversational-graph conditions within a controlled experiment, it was shown that different interface modalities reshape both performance outcomes and subjective experience.

Conversational interaction improved synthesis quality and reduced perceived workload by allowing users to refine queries iteratively and build momentum across turns. At the same time, it raised self-reported confidence without corresponding gains in calibration, highlighting the distinction between perceived fluency and actual correctness. The addition of a relational graph enabled richer cross-item reasoning but also increased cognitive effort

for some participants, emphasizing the importance of managing visual and interaction complexity.

Overall, the findings indicate that the value of conversational and hybrid interfaces in exploratory search depends not only on model capability but also on how interaction cues, initiative, and explanation are structured. Designing for calibrated confidence involves supporting reflection and transparency while avoiding premature closure and over-reliance on system authority. The study suggests that conversational interaction is a major factor shaping users' cognitive and affective engagement, beyond retrieval alone. Moreover, the results reinforce earlier chapters' emphasis on interaction scaffolding and knowledge graph mediation as design strategies for balancing system initiative and user agency in exploratory information systems.

Beyond digital libraries, these insights extend to broader human–AI collaboration scenarios where users must navigate large, semantically rich information spaces. In particular, they strengthen the thesis argument that grounding conversational interfaces in structured, explainable representations such as knowledge graphs can support interpretability and calibrated confidence even as systems grow in complexity.

Future research should investigate how such interfaces perform over longer periods of use, across different domains, and with diverse user populations to understand how trust and confidence evolve with continued exposure to conversational, visual, and hybrid search systems. For this thesis, the main value of the cross-domain study is that it reinforces the interpretation of Iteration 3's results as interaction-level phenomena, particularly the separation between reduced workload and unchanged or weakly changed calibration and judgment quality. This provides a natural bridge to the next chapter, which synthesizes design principles for exploratory AI systems across domains, drawing together the empirical and theoretical contributions established throughout the thesis.

Chapter 6

Conclusion

This thesis presented InterDev, an ontology-driven linked-data platform for managing the full lifecycle of international development evaluation evidence. InterDev represents evaluations using an ERCT-aligned knowledge graph and provides researcher-facing interfaces for browsing, faceted filtering, curation via collections, and export to familiar tabular formats, without requiring semantic web expertise. The system also supports controlled submission workflows for extending the knowledge graph, including optional LLM-assisted submission with human-in-the-loop review.

The core contribution of the thesis lies not only in the implementation of InterDev, but in its iterative, user-centred evaluation across three design iterations. The first iteration established baseline usability and identified breakdowns in filtering, collection management, export, and manual submission. The second and third iterations both focused on aspects of LLM-mediated interaction for linked data platforms. The second iteration introduced multi-criteria filtering, multi-collection workflows, improved export, and LLM-assisted submission to reduce contribution friction. The third iteration extended evaluation beyond usability to examine the epistemic effects of LLM-mediated exploration, looking at interaction, measuring workload, confidence and its calibration, and verification behaviour under different provenance conditions. Together, these studies show that while automation can substantially reduce interactional effort, it also reshapes users' judgment and verification practices, creating new design requirements for transparency, calibration, and support for reliable reasoning. Finally, a complementary cross-domain study in digital libraries was used as a bounded external-validity probe, providing additional evidence that these interaction-level trade-offs are not unique to international development, while stopping short of claiming generalisation of the InterDev system itself.

6.1 Summary of findings in relation to the research question

This thesis asked whether an ontology-driven linked-data platform, enhanced with LLM-mediated interaction, can improve the discovery, contribution, and reuse of international development evaluation evidence while preserving transparency, auditability, and appropriate user reliance.

6.1.1 What was demonstrated

The thesis demonstrated that an ontology-driven linked-data platform can make fragmented impact-evaluation evidence more usable for non-semantic-web researchers through faceted filtering, collection-based curation, and export-oriented workflows grounded in a shared ERCT-aligned knowledge graph. It also demonstrated that LLM-assisted submission can substantially reduce submission friction in submission workflows: compared with manual submission, LLM-assisted submission improved completion and reduced time-on-task, while remaining compatible with human review and pre-ingestion validation. In exploration settings, the thesis further demonstrated that conversational mediation can reduce perceived workload during evidence use.

6.1.2 What was partially demonstrated

The thesis partially demonstrated that provenance-aware and LLM-mediated interaction improves reliable evidence use. Surfaced provenance changed how participants judged outputs and supported a more critical stance toward generated answers, but this did not translate into improved task performance or stronger confidence calibration in Iteration 3. Similarly, conversational support reduced interactional effort, but did not reliably improve synthesis quality or efficiency once task difficulty was controlled. The cross-domain study provided bounded supporting evidence that these trade-offs are not unique to international development, but this should be interpreted as a probe of interaction-level feasibility rather than as a generalisation of InterDev itself.

6.1.3 What remains unresolved

The thesis does not establish that LLM mediation improves judgment quality in evidence synthesis, nor that provenance cues alone are sufficient to produce calibrated reliance. It also does not resolve how such systems perform under long-term, real-world use, where verification practices, organisational workflows, and evolving corpora may alter both usability and trust. More broadly, the work leaves open how best to support claim-level verification, synthesis scaffolding, and semantic validation at scale without reintroducing unacceptable cognitive burden.

6.1.4 Overall summary

Taken together, the thesis shows that ontology-driven linked-data platforms are a viable basis for reusable international development evidence infrastructures, and that LLM mediation can improve interactional efficiency, especially for contribution and exploratory access. However, it also shows that reducing interactional effort is not the same as improving epistemic outcomes. The main unresolved challenge is therefore not whether LLMs can be added to linked-data platforms, but how they can be integrated so that usability gains are matched by support for reliable comparison, verification, and synthesis work.

6.2 Findings in relation to the research objectives

This section summarises how each research objective was addressed, linking the work to the artefacts produced (InterDev and ERCT), the evaluation programme, and the resulting empirical findings. Across objectives, the thesis combines system-building with iterative, mixed-methods evaluation to establish both practical feasibility and design implications for LLM-mediated linked-data evidence platforms.

6.2.1 Research Objective 1: State of the art and derived requirements

RO1 was achieved through a targeted review of interfaces for knowledge graphs, LLM interaction, and KG and LLM integration, with selection criteria explicitly aligned to the research objectives and InterDev’s applied requirements. The review synthesised the literature across four design-relevant dimensions: interface design, user interaction, backend integration, and transparency and provenance. It then reframed these into three

evaluative lenses that InterDev must satisfy in practice: accessibility for non-semantic web users, interpretability for auditable and trustworthy use, and functionality sufficient for expressive retrieval and evidence synthesis. This analysis positioned InterDev against recurring limitations in existing KG front ends, including scalability and cognitive burden in graph visualisation and limits of natural language interfaces for complex queries, and against known risks in LLM interfaces, including hallucination, overreliance driven by fluency, and mixed evidence on the effectiveness of explanation cues. It also identified integration patterns such as retrieval augmented generation and iterative reasoning as technically promising but interactionally under specified.

The outcome of RO1 was a concrete requirements set grounded in the workflows of systematic evidence synthesis in international development. Drawing on the systematic review analysis, the thesis derived initial user requirements for advanced filtering and contextualisation, access to both granular and aggregated representations, and tools for cross study comparison and harmonisation, alongside the interaction principles needed to make these requirements usable in practice, including progressive disclosure, scaffolding for non-expert users, and export to familiar formats. In parallel, the KG and LLM interface literature motivated explicit design attention to provenance and trust calibration, treating transparency as an interaction requirement rather than a documentation concern. This directly shaped the later evaluation focus on workload, confidence calibration, and verification behaviour under controlled provenance conditions.

6.2.2 Research Objective 2: InterDev design, implementation, and formative evaluation

RO2 focused on designing and implementing InterDev as an ERCT-aligned linked-data platform that enables non-semantic-web researchers to browse, filter, curate, and export evaluation evidence. This objective was met through the construction of the InterDev framework as an integrated system comprising: an ERCT-aligned knowledge graph; interaction primitives for evidence browsing and faceted filtering; collection-based curation as a bounded evidence workspace; and export mechanisms that support reuse in familiar analysis workflows.

The formative evaluation programme provides evidence that the platform can be used effectively by target users while also revealing where interaction breakdowns occur.

Iteration 1 established baseline usability and identified concrete frictions in filtering, collection management, export affordances, and manual submission. These findings guided subsequent redesign and provide a validated account of the initial barriers that prevent linked-data platforms from being practical for non-expert users. Iteration 2 then demonstrated that targeted interface and workflow redesign substantially improves usability outcomes and efficiency for core tasks, supporting the claim that InterDev’s interaction model can make knowledge-graph evidence usable without requiring semantic web literacy. In this sense, RO2 is achieved not only by implementing the platform, but by empirically showing how its interaction design must be shaped to support real research workflows.

6.2.3 Research Objective 3: LLM-mediated interaction and evaluation across domains

RO3 examined whether two LLM-mediated workflows can improve both contribution and exploration in linked-data evidence systems, and how LLM-mediated interaction in these workflows affect usability and epistemic outcomes. This objective was addressed in two complementary ways.

First, the thesis evaluated LLM-assisted submission as a human-in-the-loop contribution workflow. Iteration 2 provides evidence that pre-populated submission can substantially reduce manual effort and improve completion rates relative to manual entry, while also changing users’ perceived accuracy and trust in the submission pipeline. This supports the claim that LLM-assisted submission can remove practical barriers to submission, but also motivates the need for validation mechanisms and review interfaces that preserve user oversight.

Second, the thesis evaluated LLM-mediated exploration under controlled interface conditions, evaluated on synthesis-oriented tasks. Iteration 3 extends beyond usability to measure workload, confidence and calibration, and verification behaviours when interacting with evidence through baseline browsing, chat assistance without provenance, and chat assistance with explicit provenance cues. The results show that conversational mediation can reliably reduce perceived workload, but that lower interactional effort does not necessarily yield better task performance or faster completion once task difficulty is controlled. Importantly, the provenance manipulation provides evidence that transparency

cues can change how users experience and judge their own answers, affecting confidence and calibration without necessarily increasing trust. These findings operationalise the thesis claim that LLM integration must be evaluated in epistemic terms, not only in efficiency or satisfaction terms.

Finally, a cross-domain study in digital libraries served as a bounded external-validity probe for RO3. By testing comparable interaction modalities (keyword, semantic, conversational, conversational-graph) over a different corpus, the study provided additional evidence that modality-driven trade-offs, particularly the relationship between conversational fluency, workload reduction, and confidence, are not unique to international development evidence systems. This supports the broader relevance of the interaction-level implications while avoiding overclaiming generalisation of InterDev itself.

6.2.4 Research Objective 4: ERCT refinement and ontology quality assessment

This research objective was to refine the ERCT ontology so that it could function as more than a descriptive schema, instead serving as an operational backbone for interaction, submission, exploration, and synthesis-oriented evidence work within InterDev. Across the three evaluation iterations, the ontology was progressively reshaped in response to concrete usability breakdowns and workflow demands. Early refinements focused on improving conceptual clarity, stabilising terminology, and reducing ambiguity so that the model could support predictable filtering, clearer data entry, and more consistent interpretation by both the system and its users. This marked a shift from treating the ontology as a purely backend artefact to treating it as part of the interaction contract between users and the platform.

Subsequent iterations extended this role by strengthening the ontology’s support for structured contribution, LLM-assisted submission, and evidence use. The model was expanded to better represent how data enters the system, to support controlled vocabularies for consistent filtering and reuse, and to provide a more appropriate unit of representation for comparison and synthesis across studies. Lexical coverage was also improved to better align system representations with user language, supporting search, navigation, and interpretation without changing the underlying semantics.

Finally, the refined ontology was subjected to a assessment against the FAIR principles using the FOOPS framework, providing an external, criteria-driven evaluation of its

documentation, metadata, and reuse readiness for alignment with FAIR principles. The final version of the ontology scored a near perfect FOOPS score with 14/15 review criteria fulfilled.

6.3 Design implications for linked-data evidence platforms with LLM mediation

This section distils the thesis findings into some proposed design principles for linked-data evidence systems that combine knowledge graphs with LLM-mediated exploration. The implications focus on how to preserve the interoperability and traceability benefits of linked representations while reducing interaction cost and supporting reliable, verifiable reasoning.

Principle 1: Treat interactional effort and epistemic effort as separate design targets

Interfaces should be designed and evaluated to reduce interactional effort without assuming that this improves epistemic outcomes. The results show that LLM-mediated exploration can lower perceived workload, but this does not guarantee improved performance or faster completion once task difficulty is controlled. Systems should therefore pair efficiency-oriented interaction design with features that explicitly support epistemic work, such as structured comparison views, task decomposition, and prompts that require users to externalise assumptions and evidence before committing to an interpretation.

Principle 2: Design for calibrated confidence, not for trust as a generic outcome

LLM-mediated interfaces can shift user confidence independently of accuracy, creating a risk of miscalibration driven by fluent outputs. Design should therefore target calibration as a first-class goal. Concretely, systems should support confidence elicitation at decision points, provide lightweight self-check routines, and prioritise feedback that helps users update confidence based on evidence quality rather than language quality. Where possible, interfaces should expose verification pathways that are faster than accepting the generated response, so that checking is the default low-friction action rather than an optional extra step.

Principle 3: Make provenance a verification affordance, not a decorative citation layer

Provenance cues should be presented as actionable interaction objects that users can inspect, trace, and use to validate claims, rather than as passive references appended to answers. The controlled provenance conditions show that surfacing provenance changes how users judge and report confidence and affects verification behaviour. Effective designs should allow users to open the exact triples, entities, or source passages that support each claim, and should make it clear when a claim is a synthesis, an inference, or a direct retrieval. Provenance should also communicate absence, so users can distinguish between missing data and negative evidence.

Principle 4: Constrain LLM mediation with collection-bounded grounding and explicit scope control

Conversational interaction should be grounded within bounded evidence sets that users can curate, inspect, and export. Collection-based workspaces help to make the scope of retrieval legible, reduce inadvertent drift across heterogeneous graph regions, and align the conversational layer with established synthesis workflows. Designs should allow users to set and modify the active evidence scope, show what is currently in scope, and make scope changes explicit in the dialogue. This turns the collection into both a cognitive scaffold and a governance mechanism for what the model is permitted to summarise or compare.

Principle 5: Manage graph complexity through progressive disclosure tied to tasks

Graph interfaces should expose complexity incrementally and in ways that reflect user tasks such as filtering, comparison, and synthesis. The evaluations indicate that non-expert users encounter breakdowns when graph richness is presented without sufficient scaffolding. Progressive disclosure should be paired with clear task primitives such as faceted filters, comparison panels, and exportable summaries, so that the graph becomes a navigable substrate rather than the primary visual burden. Where graph views are used, interfaces should emphasise local neighbourhoods and controlled expansion, and provide explainable shortcuts from visual nodes to structured views.

Principle 6: Treat LLM-assisted submission as a collaborative workflow with explicit responsibility boundaries

LLM-assisted submission can reduce contribution friction and increase completion, but it should be designed as a collaboration where the system proposes structure and the user verifies and finalises it. Interfaces should make model generated fields visually distinct, support rapid correction, and prioritise validation against ontology constraints before ingest. Responsibility boundaries should be explicit, particularly for fields that affect downstream retrieval and synthesis. This reduces the risk that effort savings at ingestion create latent quality problems that reappear later during analysis.

6.4 Future Work

The evaluation programme established the feasibility of ontology-driven, LLM-mediated evidence interaction, but it also exposed clear limits. Future work should therefore focus on a small number of concrete directions that follow directly from the findings of Chapters 4–5, especially where usability gains created unresolved epistemic risks.

6.4.1 Longitudinal deployment in real evidence-synthesis settings

A first priority is longer-term, in-situ evaluation. The controlled studies in this thesis were necessary for isolating interface and mediation effects, but they do not capture how InterDev would be adopted and used across extended evidence-synthesis projects. Real-world use introduces organisational routines, collaboration practices, repeated exposure to LLM assistance, and evolving corpora, all of which may alter verification behaviour, confidence calibration, and perceived utility over time.

Future work should therefore deploy InterDev in ongoing evidence-synthesis settings over longer periods. This would allow evaluation of adoption, workflow integration, and stability of use under realistic constraints. It would also make it possible to measure outcomes that are more meaningful for practice, such as time-to-synthesis, reproducibility of exported datasets, persistence of collections across sessions, and the extent to which provenance cues support sustained rather than one-off verification behaviour.

6.4.2 Claim-level verification and calibrated reasoning support

A second priority is stronger support for verification and calibrated reasoning during evidence exploration. Iteration 3 showed that surfaced provenance changed how users judged outputs, but did not by itself improve task performance or confidence calibration. This suggests that transparency alone is insufficient when users still bear a high cognitive cost in tracing, interpreting, and reconciling evidence.

Future work should therefore investigate interaction designs that reduce the effort of checking claims against underlying evidence. This includes claim-level provenance links, interfaces that distinguish direct retrieval from synthesis or inference, structured uncertainty prompts, and support for contrastive comparison across findings. Rather than treating citations as passive references, future systems should make them actionable verification tools. Evaluation should focus explicitly on whether such designs improve calibration, checking behaviour, and judgement quality, rather than only perceived transparency.

6.4.3 Semantic robustness in LLM-assisted submission

A third priority is improving the semantic reliability of LLM-assisted submission. The results from Iteration 2 showed that pre-populated submission can substantially reduce effort and improve completion, but the earlier feasibility work also showed that syntactic validity is not enough: semantically incorrect outputs can still pass through a workflow and affect downstream retrieval, comparison, and synthesis.

Future work should therefore explore more robust submission pipelines that combine LLM-assisted submission with stronger semantic controls. Promising directions include constrained generation, shape-based validation, rule-based post-checking, entity linking, and interfaces that make low-confidence or potentially inconsistent fields explicit for user review. This work should distinguish clearly between different failure modes, such as incorrect values, mislinked entities, missing relations, and ontology misuse, and evaluate their downstream consequences for evidence quality and reuse.

6.4.4 Scalable support for synthesis over bounded evidence collections

A fourth priority is improving support for synthesis once retrieval and curation are no longer the main bottlenecks. Across Iteration 3 and the cross-domain probe, a recurring pattern was that conversational interfaces reduced interactional effort but did not

substantially improve higher-order synthesis. This suggests that the next design challenge is not simply better retrieval, but better scaffolding for integrating and comparing evidence across multiple records.

Future work should therefore focus on synthesis-oriented interaction over curated collections. This may include structured comparison views, finding-level aggregation, claim–evidence alignment, contrastive summaries, and task-adaptive summarisation of subgraphs. On the systems side, this also requires scaling collection operations and graph interaction without reintroducing excessive cognitive burden. The key challenge is to support richer reasoning while preserving the bounded, inspectable evidence contexts that made InterDev’s conversational layer experimentally tractable.

6.4.5 Extending the assessment protocol to AI-supported evidence synthesis

Beyond the empirical findings themselves, this thesis contributes an assessment protocol for studying AI-mediated evidence interaction. Across Iteration 3 and the cross-domain GutGet evaluation, the research operationalises a reusable set of measures for comparing AI-assisted and non-AI interfaces on synthesis-oriented work, including workload, task performance, confidence, confidence calibration, verification behaviour, and qualitative reliance patterns. While developed in the contexts of development evidence and digital-library exploration, this protocol has potential value as a methodological template for future studies of AI-supported reviewing and analysis of academic literature, particularly where the central concern is not only efficiency but also whether AI changes how users verify, interpret, and justify conclusions.

A further direction for future work is therefore to further apply and refine this protocol in academic-review settings, for example systematic reviewing, scoping reviews, and evidence mapping, to test whether the workload, calibration, and verification effects observed here remain stable across additional corpora and tasks in academic literature analysis.

6.5 Closing Statement

This thesis argues that improving the reuse of international development evaluation evidence requires treating practical reuse as a workflow outcome built on interoperability rather than a purely representational one. InterDev demonstrates that an ontology-driven

knowledge graph, paired with researcher-facing interaction primitives for filtering, curation, and export, can make linked data practical for evidence synthesis without requiring semantic web expertise. At the same time, the evaluation shows that introducing LLM-mediated workflows changes more than efficiency. It alters confidence, verification behaviour, and the conditions under which users consider outputs trustworthy. The central message is therefore twofold: linked data can reduce fragmentation only if it is embedded in usable contribution and reuse workflows, and LLM mediation can reduce interactional effort only if systems also provide deliberate support for calibrated, verifiable reasoning through provenance, bounded grounding, and transparent user control.

Bibliography

- Abdul Latif Jameel Poverty Action Lab. (2020, January 9). Data and Code Availability Policy. *Data and Code Availability Policy*.
<https://www.povertyactionlab.org/sites/default/files/research-resources/J-PAL%20Data%20Code%20Availability%20Policy.pdf>
- Agarwal, O., Kale, M., Ge, H., Shakeri, S., & Al-Rfou, R. (2020). *Machine translation aided bilingual data-to-text generation and semantic parsing*. 125–130.
- Ahuja, A., Baird, S., Hicks, J. H., Kremer, M., & Miguel, E. (2018). *Economics of mass deworming programs*.
- Akram-Lodhi, A. H. (2020). “Follow the yellow brick road”? Structural shortcomings in randomized control trials. *World Development*, 127, 104803.
- Akyeampong, K., Andrabi, T., Banerjee, A., Banerji, R., Dynarski, S., Glennerster, R., & Grantham-McGregor, S. (2023). 2023 Cost-Effective Approaches to Improve Global Learning—What Does Recent Evidence Tell Us Are “Smart Buys” for Improving Learning in Low-and Middle-Income Countries? *London, Washington DC, New York.: FCDO, the World Bank, UNICEF, and USAID. [https://Thedocs. Worldbank. Org/En/Doc/231d98251cf326922518be0cbe306fdc-0200022023/Rel Ated/GEEAP-Report-Smart-Buys-2023-Final. Pdf](https://Thedocs.Worldbank.Org/En/Doc/231d98251cf326922518be0cbe306fdc-0200022023/RelAted/GEEAP-Report-Smart-Buys-2023-Final.Pdf)*.
- American Economic Association. (2021, July 13). Trial Data Access. *Trial Data Access*.
<https://www.socialscienceregistry.org/site/data>
- Arenas, M., Bertails, A., Prud’hommeaux, E., & Sequeda, J. (2012). A direct mapping of relational data to RDF. *W3C Recommendation*, 27, 1–11.

- Arenas, M., Grau, B. C., Kharlamov, E., Marciuška, Š., & Zheleznyakov, D. (2016). Faceted search over RDF-based knowledge graphs. *Journal of Web Semantics*, 37, 55–74.
- Athukorala, K., Głowacka, D., Jacucci, G., Oulasvirta, A., & Vreeken, J. (2016). 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.
- Avula, S., Choi, B., & Arguello, J. (2022). The effects of system initiative during conversational collaborative search. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1), 1–30.
- Baird, S., Ferreira, F. H., Özler, B., & Woolcock, M. (2013). Relative effectiveness of conditional and unconditional cash transfers for schooling outcomes in developing countries: A systematic review. *Campbell Systematic Reviews*, 9(1), 1–124.
- Baird, S., Ferreira, F. H., Özler, B., & Woolcock, M. (2014). Conditional, unconditional and everything in between: A systematic review of the effects of cash transfer programmes on schooling outcomes. *Journal of Development Effectiveness*, 6(1), 1–43.
- Bandiera, O., Fischer, G., Prat, A., & Ytsma, E. (2016). *Do women respond less to performance pay? Building evidence from multiple experiments*.
- Banerjee, A. V., Duflo, E., & Kremer, M. (2016). The influence of randomized controlled trials on development economics research and on development policy. *The State of Economics, The State of the World*, 482–488.
- Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., & Chung, W. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv Preprint arXiv:2302.04023*.

- Barrett, C. B., & Carter, M. R. (2020). Finding our balance? Revisiting the randomization revolution in development economics ten years further on. *World Development*, 127, 104789.
- Bastagli, F., Hagen-Zanker, J., Harman, L., Barca, V., Sturge, G., Schmidt, T., & Pellerano, L. (2016). Cash transfers: What does the evidence say. *A Rigorous Review of Programme Impact and the Role of Design and Implementation Features*. London: ODI, 1(7).
- Bates, M. J. (2002). The cascade of interactions in the digital library interface. *Information Processing & Management*, 38(3), 381–400.
- Batzdorfer, V., Zenk-Möltgen, W., Young, L., Katsanidou, A., Breuer, J., & Bishop, L. (2024). Between urgency and data quality: Assessing the FAIRness of data in social science research on the COVID-19 pandemic. *Research Ethics*, 20(4), 744–763.
- Bédécarrats, F., Guérin, I., Morvant-Roux, S., & Roubaud, F. (2021). Behind the scenes of science in action: A ‘replication in context’ of a randomised control trial in Morocco. *Third World Quarterly*, 42(11), 2669–2689.
- Bédécarrats, F., Guérin, I., & Roubaud, F. (2019). All that glitters is not gold. The political economy of randomized evaluations in development. *Development and Change*, 50(3), 735–762.
- Bill and Melinda Gates Foundation. (2021, January 1). *Open Access Policy*.
- Bizer, C., Heath, T., & Berners-Lee, T. (2011). Linked data: The story so far. In *Semantic services, interoperability and web applications: Emerging concepts* (pp. 205–227). IGI global.

- Boeckhout, M., Zielhuis, G. A., & Bredenoord, A. L. (2018). The FAIR guiding principles for data stewardship: Fair enough? *European Journal of Human Genetics*, 26(7), 931–936.
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., & Brunskill, E. (2021). On the opportunities and risks of foundation models. *arXiv Preprint arXiv:2108.07258*.
- Boon, M., Hoeber, O., Hoeber, L., Storie, D., & Ramshaw, V. (2024). Exploratory Search and Beyond: A Study of Complex Search Scenarios within a Public Digital Library. *Proceedings of the Association for Information Science and Technology*, 61(1), 44–55.
- Booth, A., James, M.-S., Clowes, M., & Sutton, A. (2021). *Systematic approaches to a successful literature review*.
- Boren, T., & Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication*, 43(3), 261–278.
- Brachman, M., Kunde, S., Miller, S., Fucs, A., Dempsey, S., Jabbour, J., & Geyer, W. (2025). Building Appropriate Mental Models: What Users Know and Want to Know about an Agentic AI Chatbot. *Proceedings of the 30th International Conference on Intelligent User Interfaces*, 247–264.
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101.
- Brickley, D., Guha, R. V., & McBride, B. (2014). RDF Schema 1.1. *W3C Recommendation*, 25, 2004–2014.
- Brier, G. W. (1950). The statistical theory of turbulence and the problem of diffusion in the atmosphere. *Journal of Atmospheric Sciences*, 7(4), 283–290.

- Browne, G. R. (2024). An evidence cycle framework for community development initiatives. *Community Development Journal*, 59(1), 30–48.
- Butcher, K. R., Davies, S., Crockett, A., Dewald, A., & Zheng, R. (2011). Do graphical search interfaces support effective search for and evaluation of digital library resources. *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, 315–324.
- Cameron, D. B., Mishra, A., & Brown, A. N. (2016). The growth of impact evaluation for international development: How much have we learned? *Journal of Development Effectiveness*, 8(1), 1–21.
- Campbell, F., Tricco, A. C., Munn, Z., Pollock, D., Saran, A., Sutton, A., White, H., & Khalil, H. (2023). Mapping reviews, scoping reviews, and evidence and gap maps (EGMs): The same but different—The “Big Picture” review family. *Systematic Reviews*, 12(1), 45.
- Carevic, Z., Lusky, M., van Hoek, W., & Mayr, P. (2018). Investigating exploratory search activities based on the stratagem level in digital libraries. *International Journal on Digital Libraries*, 19(2), 231–251.
- Carroll, J. J., Bizer, C., Hayes, P., & Stickler, P. (2005). Named graphs, provenance and trust. *Proceedings of the 14th International Conference on World Wide Web*, 613–622.
- Carroll, J. M., & Rosson, M. B. (1987). Paradox of the active user. In *Interfacing thought: Cognitive aspects of human-computer interaction* (pp. 80–111).
- Chaves-Fraga, D., Corcho, O., Yedro, F., Moreno, R., Olías, J., & De La Azuela, A. (2022). Systematic construction of knowledge graphs for research-performing organizations. *Information*, 13(12), 562.

- Cheng, Z., Chen, P., Song, W., Zhang, H., Li, Z., & Sun, L. (2025). An exploratory study on how ai awareness impacts human-ai design collaboration. *Proceedings of the 30th International Conference on Intelligent User Interfaces*, 157–172.
- Clarke, V., Braun, V., & Hayfield, N. (2015). Thematic analysis. *Qualitative Psychology: A Practical Guide to Research Methods*, 3, 222–248.
- Concern Worldwide. (2022, November 12). What is International Development? *What Is International Development*. <https://www.concern.org.uk/what-is-international-development>
- Conn, K. M. (2017). Identifying effective education interventions in sub-Saharan Africa: A meta-analysis of impact evaluations. *Review of Educational Research*, 87(5), 863–898.
- Cooke, A., Smith, D., & Booth, A. (2012). Beyond PICO: the SPIDER tool for qualitative evidence synthesis. *Qualitative Health Research*, 22(10), 1435–1443.
- Dadzie, A.-S., Rowe, M., & Petrelli, D. (2011). Hide the stack: Toward usable linked data. *Extended Semantic Web Conference*, 93–107.
- De Mel, S., McKenzie, D. J., & Woodruff, C. (2009). Measuring microenterprise profits: Must we ask how the sausage is made? *Journal of Development Economics*, 88(1), 19–31.
- de Walque, D., Fernald, L., Gertler, P., & Hidrobo, M. (2017). *Cash transfers and child and adolescent development*.
- Dekker, R. (2020). Social data: CESSDA best practices. *Data Intelligence*, 2(1–2), 220–229.
- Ding, Y., Facciani, M., Joyce, E., Poudel, A., Bhattacharya, S., Veeramani, B., Aguinaga, S., & Weninger, T. (2025). Citations and trust in llm generated responses.

- Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22), 23787–23795.
- DIS, I. (2009). 9241-210: 2010. Ergonomics of human system interaction-Part 210: Human-centred design for interactive systems. *International Standardization Organization (ISO)*. Switzerland.
- Dognin, P. L., Padhi, I., Melnyk, I., & Das, P. (2021). ReGen: Reinforcement learning for text and knowledge base generation using pretrained language models. *arXiv Preprint arXiv:2108.12472*.
- Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., & Zhang, W. (2014). *Knowledge vault: A web-scale approach to probabilistic knowledge fusion*. 601–610.
- Dubey, M., Banerjee, D., Abdelkawi, A., & Lehmann, J. (2019). *Lc-quad 2.0: A large dataset for complex question answering over wikidata and dbpedia*. 69–78.
- DuCharme, B. (2013). *Learning SPARQL: querying and updating with SPARQL 1.1*. O'Reilly Media, Inc.
- Dudáš, M., Lohmann, S., Svátek, V., & Pavlov, D. (2018). Ontology visualization methods and tools: A survey of the state of the art. *The Knowledge Engineering Review*, 33, e10.
- Egloff, M., Adamou, A., & Picca, D. (2020). Enabling Ontology-Based Data Access to Project Gutenberg. *WHiSe@ ESWC*, 21–32.
- Elbers, C., & Gunning, J. W. (2014). Evaluation of development programs: Randomized controlled trials or regressions? *The World Bank Economic Review*, 28(3), 432–445.
- Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, 87(3), 215.

- Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity*, 5(3), 178–186.
- Evans, D. K., & Popova, A. (2016). What really works to improve learning in developing countries? An analysis of divergent findings in systematic reviews. *The World Bank Research Observer*, 31(2), 242–270.
- Ferré, S. (2016). Sparklis: An expressive query builder for SPARQL endpoints with guidance in natural language. *Semantic Web*, 8(3), 405–418.
- Francart, T. (2023). *Sparnatural: A visual knowledge graph exploration tool*. 11–15.
- Fryer Jr, R. G. (2017). The production of human capital in developed countries: Evidence from 196 randomized field experiments. In *Handbook of economic field experiments* (Vol. 2, pp. 95–322). Elsevier.
- Ganimian, A. J., & Murnane, R. J. (2016). Improving education in developing countries: Lessons from rigorous impact evaluations. *Review of Educational Research*, 86(3), 719–755.
- Gareth B Johnson. (2025). *Gutendex: Web API for Project Gutenberg ebook metadata*. <https://github.com/garethbjohnson/gutendex>
- Garijo, D., Corcho, O., & Poveda-Villalón, M. (2021). *FOOPS!: An Ontology Pitfall Scanner for the FAIR principles*. ISWC (Posters/Demos/Industry).
- Gerlach, M., & Font-Clos, F. (2020). A standardized project gutenber corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*, 22(1), 126.
- Gero, K. I., Swoopes, C., Gu, Z., Kummerfeld, J. K., & Glassman, E. L. (2024). Supporting sensemaking of large language model outputs at scale. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–21.

- Gertler, P., Galiani, S., & Romero, M. (2018). *How to make replication the norm*.
- Ghoniem, M., Fekete, J.-D., & Castagliola, P. (2004). *A comparison of the readability of graphs using node-link and matrix-based representations*. 17–24.
- Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30), e2305016120.
- Glennerster, R., & Takavarasha, K. (2014). *Running randomized evaluations: A practical guide*. Princeton University Press.
- Glewwe, P. W., Hanushek, E. A., Humpage, S. D., & Ravina, R. (2011). *School resources and educational outcomes in developing countries: A review of the literature from 1990 to 2010*.
- Grant, S., & Khatua, S. (2024). Research transparency and reproducibility policies and programmes at the international initiative for impact evaluation. *Journal of Development Effectiveness*, 16(3), 363–373.
- Greenhalgh, T., & Peacock, R. (2005). Effectiveness and efficiency of search methods in systematic reviews of complex evidence: Audit of primary sources. *Bmj*, 331(7524), 1064–1065.
- Groth, P., Gibson, A., & Velterop, J. (2010). The anatomy of a nanopublication. *Information Services and Use*, 30(1–2), 51–56.
- Harris, S. (2013). SPARQL 1.1 Query Language. *W3C Recommendation*, 21.
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139–183). Elsevier.
- Harvard College. (2023, February 13). Harvard Dataverse. *Harvard Dataverse*.
- Hearst, M. (2009). *Search user interfaces*. Cambridge university press.

- Heath, T., & Bizer, C. (2011). *Linked data: Evolving the web into a global data space*. Morgan & Claypool Publishers.
- Heeks, R. (2002). Failure, success and improvisation of information systems projects in developing countries. *Development Informatics Working Paper*, (11).
- Hodrien, A., & Fernando, T. (2021). A review of post-study and post-task subjective questionnaires to guide assessment of system usability. *Journal of Usability Studies*, 16(3).
- Hoerber, O., & Shukla, S. (2022). A study of visually linked keywords to support exploratory browsing in academic search. *Journal of the Association for Information Science and Technology*, 73(8), 1171–1191.
- Hoefler, P. (2013). *Linked data interfaces for non-expert users*. 702–706.
- Hughes, L. D., Tsueng, G., DiGiovanna, J., Horvath, T. D., Rasmussen, L. V., Savidge, T. C., Stoeger, T., Turkarslan, S., Wu, Q., Wu, C., & others. (2023). Addressing barriers in FAIR data practices for biomedical data. *Scientific Data*, 10(1), 98.
- International Initiative for Impact Evaluation. (2022, April 4). *3ie Development Evidence Portal*. <https://www.3ieimpact.org/evidence-hub>
- International Initiative for Impact Evaluation 2021). (2021, July 13). About RIDIE. *About RIDIE*. <https://ridie.3ieimpact.org/index.php?r=site/page&view=about>
- ISO/TC 159/SC 4 Ergonomics of Human-System Interaction (Subcommittee). (1998). *Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs): Guidance on Usability*. International Organization for Standardization.
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38.

- Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1), 53–71.
- Jiang, M., Dubnicek, R. C., Worthey, G., Underwood, T., & Downie, J. S. (2022). A prototype Gutenberg-Hathitrust sentence-level parallel corpus for OCR error analysis: Pilot investigations. *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, 1–5.
- Jiang, P., Rayan, J., Dow, S. P., & Xia, H. (2023). Graphologue: Exploring large language model responses with interactive diagrams. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–20.
- Kabeer, N. (2020). Women’s empowerment and economic development: A feminist critique of storytelling practices in “randomista” economics. *Feminist Economics*, 26(2), 1–26.
- Kamdar, M. R., & Musen, M. A. (2021). An empirical meta-analysis of the life sciences linked open data on the web. *Scientific Data*, 8(1), 24.
- Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C., & Giannopoulou, E. (2007). Ontology visualization methods—A survey. *ACM Computing Surveys (CSUR)*, 39(4), 10-es.
- Klyne, G. (2004). RDF Concepts and Abstract Syntax W3C Recommendation. [Http://Www. W3. Org/TR/Rdf-Concepts/](http://www.w3.org/TR/Rdf-Concepts/).
- Kopper, S., Sautmann, A., & Turitto, J. (2020, January 1). Data Publication. *Data Publication*. <https://www.povertyactionlab.org/resource/data-publication>
- Kremer, M., Brannen, C., & Glennerster, R. (2013). The challenge of education and learning in the developing world. *Science*, 340(6130), 297–300.
- Kriglstein, S. (2009). *User requirements analysis on ontology visualization*. 694–699.

- Krishnaratne, S., & White, H. (2013). *Quality education for all children? What works in education in developing countries*. International Initiative for Impact Evaluation (3ie).
- Kugler, A. D., & Rojas, I. (2018). *Do CCTs improve employment and earnings in the very long-term? Evidence from Mexico*. National Bureau of Economic Research.
- Kuhn, T., & Dumontier, M. (2014). Trusty URIs: Verifiable, immutable, and permanent digital artifacts for linked data. *European Semantic Web Conference*, 395–410.
- Kujala, S. (2003). User involvement: A review of the benefits and challenges. *Behaviour & Information Technology*, 22(1), 1–16.
- Kuric, E., Fernández, J. D., & Drozd, O. (2019). *Knowledge graph exploration: A usability evaluation of query builders for laypeople*. 326–342.
- Lagarde, M., Haines, A., & Palmer, N. (2007). Conditional cash transfers for improving uptake of health interventions in low-and middle-income countries: A systematic review. *Jama*, 298(16), 1900–1910.
- Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., & Zhao, J. (2013). *Prov-o: The prov ontology*.
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80.
- Lewis, J. R. (1992). *Psychometric evaluation of the post-study system usability questionnaire: The PSSUQ*. 36(16), 1259–1260.
- Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, 7(1), 57–78.

- Lewis, J. R. (2002). Psychometric evaluation of the PSSUQ using data from five years of usability studies. *International Journal of Human-Computer Interaction*, 14(3–4), 463–488.
- Lewis, J. R. (2012). Usability testing. *Handbook of Human Factors and Ergonomics*, 1267–1312.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., & Rocktäschel, T. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
- Li, J., Yang, Y., Liao, Q. V., Zhang, J., & Lee, Y.-C. (2025). As Confidence Aligns: Understanding the Effect of AI Confidence on Human Self-confidence in Human-AI Decision Making. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–16.
- Li, X., Wang, Y., Wang, H., Wang, Y., & Zhao, J. (2021). Nbsearch: Semantic search and visual exploration of computational notebooks. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14.
- Liang, S., Stockinger, K., De Farias, T. M., Anisimova, M., & Gil, M. (2021). Querying knowledge graphs in natural language. *Journal of Big Data*, 8(1), 3.
- Lin, C.-Y. (2004). *Rouge: A package for automatic evaluation of summaries*. 74–81.
- Linckels, S., & Meinel, C. (2011). *E-librarian service: User-friendly semantic search in digital libraries*. Springer.
- Liu, Y., Li, S., Huang, K., & Wang, Q. (2024). *AutoIE: An Automated Framework for Information Extraction from Scientific Literature*. 424–436.
- Ma, S., Wang, X., Lei, Y., Shi, C., Yin, M., & Ma, X. (2024a). “Are you really sure?” Understanding the effects of human self-confidence calibration in AI-assisted

- decision making. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–20.
- Ma, S., Wang, X., Lei, Y., Shi, C., Yin, M., & Ma, X. (2024b). “Are you really sure?” Understanding the effects of human self-confidence calibration in AI-assisted decision making. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–20.
- MacKenzie, I. S. (2024). *Human-computer interaction: An empirical research perspective*.
- Malle, B. F., & Ullman, D. (2021). A multidimensional conception and measure of human-robot trust. In *Trust in human-robot interaction* (pp. 3–25). Elsevier.
- Mallett, R., Hagen-Zanker, J., Slater, R., & Duvendack, M. (2012). The benefits and challenges of using systematic reviews in international development research. *Journal of Development Effectiveness*, 4(3), 445–455.
- Manghi, P. (2024). Challenges in building scholarly knowledge graphs for research assessment in open science. *Quantitative Science Studies*, 5(4), 991–1021.
- Marchionini, G. (2006). Exploratory search: From finding to understanding. *Communications of the ACM*, 49(4), 41–46.
- Markowitz, E., Ramakrishna, A., Dhamala, J., Mehrabi, N., Peris, C., Gupta, R., Chang, K.-W., & Galstyan, A. (2024). Tree-of-traversals: A zero-shot reasoning algorithm for augmenting black-box language models with knowledge graphs. *arXiv Preprint arXiv:2407.21358*.
- Marshall, I. J., & Wallace, B. C. (2019). Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. *Systematic Reviews*, 8(1), 163.

- Martorana, M., Kuhn, T., Siebes, R., & van Ossenbruggen, J. (2022). Aligning restricted access data with FAIR: a systematic review. *PeerJ Computer Science*, 8, e1038.
- Masino, S., & Niño-Zarazúa, M. (2016). What works to improve the quality of student learning in developing countries? *International Journal of Educational Development*, 48, 53–65.
- Matsumoto, N., Moran, J., Choi, H., Hernandez, M. E., Venkatesan, M., Wang, P., & Moore, J. H. (2024). KRAGEN: a knowledge graph-enhanced RAG framework for biomedical problem solving using large language models. *Bioinformatics*, 40(6), btae353.
- Matt Murtagh-White, Declan O’Sullivan, & PJ Wall. (2023). Learning from the Evidence: Impact Evaluations, Ontology and Policy. *2023 IEEE 17th International Conference on Semantic Computing (ICSC)*.
- McEwan, P. J. (2015). Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments. *Review of Educational Research*, 85(3), 353–394.
- McGuinness, D. L., & Van Harmelen, F. (2004). OWL web ontology language overview. *W3C Recommendation*, 10(10), 2004.
- McGuire, J., Kaiser, C., & Bach-Mortensen, A. M. (2022). A systematic review and meta-analysis of the impact of cash transfers on subjective well-being and mental health in low-and middle-income countries. *Nature Human Behaviour*, 6(3), 359–370.
- Meager, R. (2019). Understanding the average impact of microcredit expansions: A bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, 11(1), 57–91.

- Melnyk, I., Dognin, P., & Das, P. (2022). Knowledge graph generation from text. *arXiv Preprint arXiv:2211.10511*.
- Menon, P., Ruel, M. T., Nguyen, P. H., Kim, S. S., Lapping, K., Frongillo, E. A., & Alayon, S. (2020). Lessons from using cluster-randomized evaluations to build evidence on large-scale nutrition behavior change interventions. *World Development*, 127, 104816.
- Methley, A. M., Campbell, S., Chew-Graham, C., McNally, R., & Cheraghi-Sohi, S. (2014). PICO, PICOS and SPIDER: a comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews. *BMC Health Services Research*, 14(1), 1–10.
- Metzger, L., Miller, L., Baumann, M., & Kraus, J. (2024). Empowering calibrated (dis-) trust in conversational agents: A user study on the persuasive power of limitation disclaimers vs. Authoritative style. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–19.
- Metzger, M. J. (2007). Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology*, 58(13), 2078–2091.
- Michaelowa, A., & Michaelowa, K. (2011). Coding error or statistical embellishment? The political economy of reporting climate aid. *World Development*, 39(11), 2010–2020.
- Miguel, E. (2021). Evidence on research transparency in economics. *Journal of Economic Perspectives*, 35(3), 193–214.
- Miguel, E., & Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, 72(1), 159–217.

- Mihindukulasooriya, N., Tiwari, S., Enguix, C. F., & Lata, K. (2023). *Text2kgbench: A benchmark for ontology-driven knowledge graph generation from text*. 247–265.
- Millán, T. M., Barham, T., Macours, K., Maluccio, J. A., & Stampini, M. (2019). Long-term impacts of conditional cash transfers: Review of the evidence. *The World Bank Research Observer*, 34(1), 119–159.
- Mo, F., Mao, K., Zhao, Z., Qian, H., Chen, H., Cheng, Y., Li, X., Zhu, Y., Dou, Z., & Nie, J.-Y. (2025). A survey of conversational search. *ACM Transactions on Information Systems*, 43(6), 1–50.
- Murtagh-White, M. (2021). *ERCT: An Ontology for Describing Randomised Controlled Trials in the Social Sciences*.
- Murtagh-White, M., Wall, P., & O’Sullivan, D. (2024). *Evaluating InterDev: A FAIR Platform for International Development Data*.
- Nielsen, J. (1994). *Usability engineering*. Morgan Kaufmann.
- Norman, D. A. (1988). *The psychology of everyday things*.
- Nourani, M., Kabir, S., Mohseni, S., & Ragan, E. D. (2019). *The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems*. 7, 97–105.
- Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods*, 16(1), 1609406917733847.
- Obrezkov, D., Oelen, A., & Auer, S. (2023). *An Approach to Evaluate User Interfaces in a Scholarly Knowledge Communication Domain*. 408–412.
- Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—A web and mobile app for systematic reviews. *Systematic Reviews*, 5(1), 210.

- Ozier, O. (2021). Replication Redux: The reproducibility crisis and the case of deworming. *The World Bank Research Observer*, 36(1), 101–130.
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., & Brennan, S. E. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Bmj*, 372.
- Palm, R. B., Laws, F., & Winther, O. (2019). *Attend, copy, parse end-to-end information extraction from documents*. 329–336.
- Papenmeier, A., Kern, D., Hienert, D., Sliwa, A., Aker, A., & Fuhr, N. (2021). Starting conversations with search engines-interfaces that elicit natural language queries. *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, 261–265.
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). *Bleu: A method for automatic evaluation of machine translation*. 311–318.
- Parker, S. W., & Vogl, T. (2018). *Do conditional cash transfers improve economic outcomes in the next generation? Evidence from Mexico*. National Bureau of Economic Research.
- Pearce, J., Chang, S., Alzougool, B., Kennedy, G., Ainley, M., & Rodrigues, S. (2011). Search or explore: Do you know what you're looking for? *Proceedings of the 23rd Australian Computer-Human Interaction Conference*, 253–256.
- Pega, F., Pabayo, R., Benny, C., Lee, E.-Y., Lhachimi, S. K., & Liu, S. Y. (2022). Unconditional cash transfers for reducing poverty and vulnerabilities: Effect on use of health services and health outcomes in low-and middle-income countries. *Cochrane Database of Systematic Reviews*, (3).

- Pirolli, P., & Card, S. (2005). The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. *Proceedings of International Conference on Intelligence Analysis*, 5(1), 2–4.
- Poveda-Villalón, M., Suárez-Figueroa, M. C., & Gómez-Pérez, A. (2012). Validating ontologies with oops! *International Conference on Knowledge Engineering and Knowledge Management*, 267–281.
- Pramanik, S., Alabi, J., Roy, R. S., & Weikum, G. (2024). UNIQORN: unified question answering over RDF knowledge graphs and natural language text. *Journal of Web Semantics*, 83, 100833.
- Preece, J., Rogers, Y., & Sharp, H. (2015). *Interaction Design: Beyond Human-Computer Interaction* (4th ed.). John Wiley & Sons.
- Project Gutenberg Literary Archive Foundation. (1971). *Project Gutenberg*.
<https://www.gutenberg.org/>
- Quisumbing, A. R., Ahmed, A., Gilligan, D. O., Hoddinott, J., Kumar, N., Leroy, J. L., Menon, P., Olney, D. K., Roy, S., & Ruel, M. (2020). Randomized controlled trials of multi-sectoral programs: Lessons from development research. *World Development*, 127, 104822.
- Reichert, L., Rogers, Y., Capra, L., Wood, E., Duong, T. D., & Sebire, N. (2022). It's good to talk: A comparison of using voice versus screen-based interactions for agent-assisted tasks. *ACM Transactions on Computer-Human Interaction*, 29(3), 1–41.
- Reiley, D. H., & List, J. A. (2007). Field experiments in economics. *The New Palgrave Dictionary of Economics*.

- Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., & Koffel, J. B. (2021). PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. *Systematic Reviews*, 10(1), 39.
- Reynolds, L., & McDonell, K. (2021). *Prompt programming for large language models: Beyond the few-shot paradigm*. 1–7.
- Richardson, W. S., Wilson, M. C., Nishikawa, J., & Hayward, R. S. (1995). The well-built clinical question: A key to evidence-based decisions. *ACP Journal Club*, 123(3), A12-3.
- Riddell, A. B. (2022). Reliable editions from unreliable components: Estimating ebooks from print editions using profile hidden markov models. *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, 1–5.
- Robertson, J., & Kaptein, M. (2016). An introduction to modern statistical methods in HCI. In *Modern Statistical Methods for HCI* (pp. 1–14). Springer.
- Rodriguez-Segura, D. (2022). EdTech in developing countries: A review of the evidence. *The World Bank Research Observer*, 37(2), 171–203.
- Rosen, R. K., Gainey, M., Nasrin, S., Garbern, S. C., Lantini, R., Elshabassi, N., Sultana, S., Hasnin, T., Alam, N. H., & Nelson, E. J. (2023). Use of Framework Matrix and Thematic Coding Methods in Qualitative Analysis for mHealth: The FluidCalc app. *International Journal of Qualitative Methods*, 22, 16094069231184123.
- Sachs, J. D. (2012). From millennium development goals to sustainable development goals. *The Lancet*, 379(9832), 2206–2211.
- Saffer, D. (2010). *Designing for interaction: Creating innovative applications and devices*. New Riders.

- Salman, M., Haller, A., Méndez, S. J. R., & Naseem, U. (2024). Doc-KG: Unstructured documents to knowledge graph construction, identification and validation with Wikidata. *Expert Systems*, e13617.
- Sauro, J., & Lewis, J. R. (2010). Average task times in usability tests: What to report? *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2347–2350.
- Schultz, T. P. (2004). School subsidies for the poor: Evaluating the Mexican Progresa poverty program. *Journal of Development Economics*, 74(1), 199–250.
- Setlur, V., Kanyuka, A., & Srinivasan, A. (2023). Olio: A semantic search interface for data repositories. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–16.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin.
- Shamseer, L., Moher, D., Clarke, M., Gherzi, D., Liberati, A., Petticrew, M., Shekelle, P., & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: Elaboration and explanation. *Bmj*, 349.
- Shanahan, H., & Bezuidenhout, L. (2022). Rethinking the A in FAIR data: Issues of data access and accessibility in research. *Frontiers in Research Metrics and Analytics*, 7, 912456.
- Simhi, A., Itzhak, I., Barez, F., Stanovsky, G., & Belinkov, Y. (2025). Trust Me, I’m Wrong: High-Certainty Hallucinations in LLMs. *arXiv Preprint arXiv:2502.12964*.
- Snilstveit, B., Stevenson, J., Phillips, D., Vojtkova, M., Gallagher, E., Schmidt, T., Jobse, H., Geelen, M., Pastorello, M. G., & Eysers, J. (2015). Interventions for improving

- learning outcomes and access to education in low-and middle-income countries: A systematic review. *3ie Systematic Review*, 24.
- Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437.
- Staab, S., & Studer, R. (2013). *Handbook on ontologies*. Springer Science & Business Media.
- Staworko, S., Boneva, I., Gayo, J. E. L., Hym, S., Prud’Hommeaux, E. G., & Solbrig, H. (2015). Complexity and Expressiveness of ShEx for RDF. *18th International Conference on Database Theory (ICDT 2015)*.
- Stewart, R. (2021). Turning ‘evidence for development’ on its head: A view from Africa. *Research for All*, 5(1).
- Stewart, R. (2023). How evidence, implementation, policy, and politics come together within evidence systems: Lessons from South Africa. *Development Policy Review*, 41(2), e12657.
- Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., Mayer, L. W., & Smyth, P. (2025). What large language models know and what people think they know. *Nature Machine Intelligence*, 7(2), 221–231.
- Stöterau, J., Kemper, J., & Ghisletta, A. (2022). The impact of vocational training interventions on youth labor market outcomes: A meta-analysis. *Available at SSRN 4217580*.
- Sun, K., Qiu, Z., Salinas, A., Huang, Y., Lee, D.-H., Benjamin, D., Morstatter, F., Ren, X., Lerman, K., & Pujara, J. (2022). Assessing scientific research papers with knowledge graphs. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2467–2472.

- Tripney, J. S., & Hombrados, J. G. (2013). Technical and vocational education and training (TVET) for young people in low-and middle-income countries: A systematic review and meta-analysis. *Empirical Research in Vocational Education and Training*, 5, 1–14.
- Tsafnat, G., Glasziou, P., Choong, M. K., Dunn, A., Galgani, F., & Coiera, E. (2014). Systematic review automation technologies. *Systematic Reviews*, 3(1), 74.
- Vaessen, J., Lemire, S., & Befani, B. (2020). *Evaluation of international development interventions: An overview of approaches and methods*.
- Van De Schoot, R., De Bruin, J., Schram, R., Zahedi, P., De Boer, J., Weijdem, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., & others. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, 3(2), 125–133.
- Van Hooland, S., & Verborgh, R. (2014). *Linked Data for Libraries, Archives and Museums: How to clean, link and publish your metadata*. Facet publishing.
- Van Rooyen, C., Stewart, R., & De Wet, T. (2012). The impact of microfinance in sub-Saharan Africa: A systematic review of the evidence. *World Development*, 40(11), 2249–2262.
- Verborgh, R., Vander Sande, M., Hartig, O., Van Herwegen, J., De Vocht, L., De Meester, B., Haesendonck, G., & Colpaert, P. (2016). Triple pattern fragments: A low-cost knowledge graph interface for the web. *Journal of Web Semantics*, 37, 184–206.
- Waddington, H., White, H., Snilstveit, B., Hombrados, J. G., Vojtkova, M., Davies, P., Bhavsar, A., Eyers, J., Koehlmoos, T. P., & Petticrew, M. (2012). How to do a good systematic review of effects in international development: A tool kit. *Journal of Development Effectiveness*, 4(3), 359–387.

- Wang, L., Friedman, N., Zhu, C., Zhu, Z., & Mountford, S. J. (2025). The Impact of Confidence Ratings on User Trust in Large Language Models. *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, 365–370.
- Wang, X., Wang, X., Park, S., & Yao, Y. (2025). Mental Models of Generative AI Chatbot Ecosystems. *Proceedings of the 30th International Conference on Intelligent User Interfaces*, 1016–1031.
- White, R. W. (2016). *Interactions with Search Systems*. Cambridge University Press.
- White, R. W. (2024). Advancing the search frontier with AI agents. *Communications of the ACM*, 67(9), 54–65.
- Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., & Bourne, P. E. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), 1–9.
- Wood, B. D. K., Müller, R., & Brown, A. N. (2018). Push button replication: Is impact evaluation evidence for international development verifiable? *PLoS ONE*, 13(12), 1–15. Academic Search Complete.
- Xu, Z., Song, T., & Lee, Y.-C. (2025). Confronting verbalized uncertainty: Understanding how LLM’s verbalized uncertainty influences users in AI-assisted decision-making. *International Journal of Human-Computer Studies*, 197, 103455.
- Yasunaga, M., Ren, H., Bosselut, A., Liang, P., & Leskovec, J. (2021). QA-GNN: Reasoning with language models and knowledge graphs for question answering. *arXiv Preprint arXiv:2104.06378*.

- Ye, Y., Ren, J., Wang, S., Wan, Y., Razzak, I., Hoex, B., Wang, H., Xie, T., & Zhang, W. (2024). Construction and Application of Materials Knowledge Graph in Multidisciplinary Materials Science via Large Language Model. *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*.
- Yun, H. S., & Bickmore, T. (2025). Framing Health Information: The Impact of Search Methods and Source Types on User Trust and Satisfaction in the Age of LLMs. *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–7.
- Zerilli, J., Bhatt, U., & Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns*, 3(4).
- Zinke-Wehlmann, C., Kirschenbaum, A., Palma, R., Brahma, S., Charvát, K., Charvát Jr, K., & Reznik, T. (2021). Linked Data and Metadata. *Big Data in Bioeconomy: Results from the European DataBio Project*, 79–90.

Appendix

A. InterDev Study Iteration 3 Protocol

This appendix contains the full facilitator and participant protocol used in the Iteration 3 evaluation of InterDev. It documents the study flow, condition sequencing, task structure, briefing scripts, and post-condition measures used in the within-subjects comparison of baseline browsing, LLM-mediated exploration, and LLM-mediated exploration with surfaced citations. It is included to support methodological transparency and reproducibility for the evaluation reported in Section 5.6.

InterDev

Purpose

To evaluate how three different interface designs (Baseline Search, LLM Chat, LLM Chat + Knowledge Graph Citations) support users in retrieving, interpreting, and synthesizing international development evidence within curated collections. We measure task performance, verification behaviour, reliance, workload, subjective impressions, trust, and perceived accuracy across interfaces.

Setting

The study will be conducted online inside the Qualtrics survey platform, with external task execution running in the [InterDev](#) system. Participants complete the study remotely on their own devices. The study session is expected to take approximately 40 minutes, though timing may vary depending on the user's pace. All instructions, task transitions, and subjective assessments are delivered via Qualtrics, while retrieval/synthesis/validation tasks occur within embedded links to InterDev.

Central Research Question

Below is the central research question for this thesis. Please keep this in mind throughout the study.

“How does adding explicit source/citation surfacing (via a knowledge graph) to an LLM chat interface affect users’ ability to retrieve, synthesize, and verify development evidence, and how does it change trust, perceived accuracy, workload, and reliance compared to chat without sources?”

Flow Overview

Each participant will use all three interface variants, but the order in which they encounter them will be counterbalanced using a Latin Square design. This ensures that each condition appears equally often in the first, second, and third block positions across participants, mitigating order effects. In addition, task order is balanced independently, such that each of the three task-pairs appears in each block position across the study sample.

- **Sequence A:** C0 → C1 → C2
- **Sequence B:** C1 → C2 → C0
- **Sequence C:** C2 → C0 → C1

The balancing and distributing of conditions across participants is entirely handled by Qualtrics, you don't need to think about it.

Introduction & Demographics (6 questions)

Participants begin with a short introduction outlining the purpose and structure of the study. This is followed by demographic and background questions capturing domain exposure, familiarity with large language models, and related attributes.

1. Interface Sequence

Participants then complete three evaluation blocks. Each block corresponds to one of the following interface conditions:

Each interface sequence consists of:

- **Tutorial Task (1 example task preceded by a short demo by the facilitator):** A low-stakes practice task to allow participants to become familiar with the interface before the main evaluation.
- **Task Notification:** A reminder that the upcoming tasks are part of the formal evaluation and should be attempted carefully.
- **2 Experimental Tasks:** Participants complete two tasks within the same policy domain theme (fixed ordering inside the pair). Each pair is unique per block, ensuring all participants complete all six tasks across the study, but never repeat a task.
- **NASA-TLX Workload Assessment (6 sliders):** After each block, participants complete a short assessment using NASA-TLX and brief trust / verification questions.
- **Trust Scale:** After each block, participants complete a short assessment.

2. Post Evaluation Discussion and PSSUQ

Following the completion of all conditions, the participant completes a **Post Study System Usability Questionnaire (PSSUQ)**. This is a 19 question Likert scale measure. It is built into Qualtrics, the participant can complete it themselves.

Following this participants participate in a short interview about their experiences with the interfaces. This includes preference, perceived trustworthiness, task-interface fit, perceived accuracy, and how the presence or absence of surfaced citations influenced verification behaviour. The questions will appear on Qualtrics, you will be the one to ask the questions and guide the conversation.

Facilitator Instructions – Study Introduction

This study is run synchronously over Zoom with live facilitation. Before beginning, greet the participant, confirm their audio and video connection are working, and explain that you will guide them through each stage of the study. Participants will complete all tasks on their own machine, in their own browser environment.

Provide the participant with the Qualtrics survey link:

https://scsstcd.qualtrics.com/jfe/form/SV_dgPQVcVNYiL6QOG

Ask them to open the link in a modern desktop browser (Chrome recommended) and maximise the browser window.

Participants will be asked to complete the consent form before starting the survey. They will be asked to upload their completed consent form here before continuing. Participants should complete the Qualtrics sequence independently; the facilitator is present to clarify instructions and procedural questions, but not to intervene in the task execution.

Participants will be required to navigate to InterDev themselves during the study (when instructed by the survey).

They should open a second tab for:

<https://interdev.adaptcentre.ie/>

The participant will switch between Qualtrics and InterDev as needed. The facilitator should confirm that the participant is able to access InterDev before beginning the first tutorial.

Participants should share their screen during the session. This enables the facilitator to observe how they are interacting with the system for research validity purposes (e.g., where they navigate, what verification behaviour they exhibit, if clarification is needed). If the user accidentally refreshes or closes the Qualtrics tab, it will automatically reopen where you left off if you return to the survey via the link.

Example Introduction

“Hello, and thank you for joining. Before we begin, I’d like to briefly explain how this session will run. You will be completing this study on your own computer, and I will be here the entire time to guide the process, answer questions, and provide clarification when needed.”

“You’ll have two browser tabs open during the session: the Qualtrics survey, which will guide the experimental flow, and the InterDev system, which you will use to carry out the tasks. The survey will instruct you when to switch between the two.”

“Please open the Qualtrics survey now. You can upload your completed consent form onto this page. If you haven’t yet filled it out please download and do so now. Participation is voluntary and you can stop at any time. Your data belongs to you: if you want it to be deleted at any time you simply need to let us know”

“In this study, you’ll be interacting with three different versions of an evidence synthesis interface for international development and policy research. Each interface offers a different type of assistance. For each one, you’ll complete a short tutorial, then two evaluation tasks, and then a short questionnaire about your experience using that interface. The session should take approximately 40 minutes in total.”

“If anything is unclear at any point, just let me know. You can ask clarifying questions at any time.”

“Before we start, can you please confirm that you can see the Qualtrics survey clearly on your screen, and that you are able to also open a second tab for interdev.adaptcentre.ie when prompted?”

Pre-Study Questionnaire

At the beginning of the Qualtrics survey, participants will be presented with a short questionnaire that collects background information. They will be asked to provide a pseudonym to identify them for the purposes of the study. This includes questions about their familiarity with international development evidence (e.g. reading reports, working

with evaluations), their familiarity with large language models, and their awareness of online repositories or platforms that provide development research or evidence summaries. This is followed by basic demographic information such as age, gender, and first language. Introduce this section by saying:

“Before we begin the main part of the study, the survey will ask you a few short questions about your background, including your familiarity with development research and use of AI tools, as well as some demographic information. These questions help us contextualise the results and understand differences between people’s prior knowledge and experience. These are standard research questions and will only take a few minutes. Please answer them honestly. You may select ‘Prefer not to say’ for any question where you would rather not provide a response. Once you complete this section, the survey will automatically guide you to the first interface tutorial. You will also be asked to choose a pseudonym to identify you for the purposes of the study, which we ask you to use again when logging into the system later”

Ensure that the participant understands that these questions are for research context only, that their responses are confidential, and that they may skip any question they do not wish to answer (e.g., “Prefer not to say” options).

NB Once the Pre-Study Questionnaire is complete please advise the participant that you will be now turning on the recording on Zoom and ensure that they are aware and comfortable with this. Then don’t forget to turn it on!

Facilitator Pre-Session Checklist

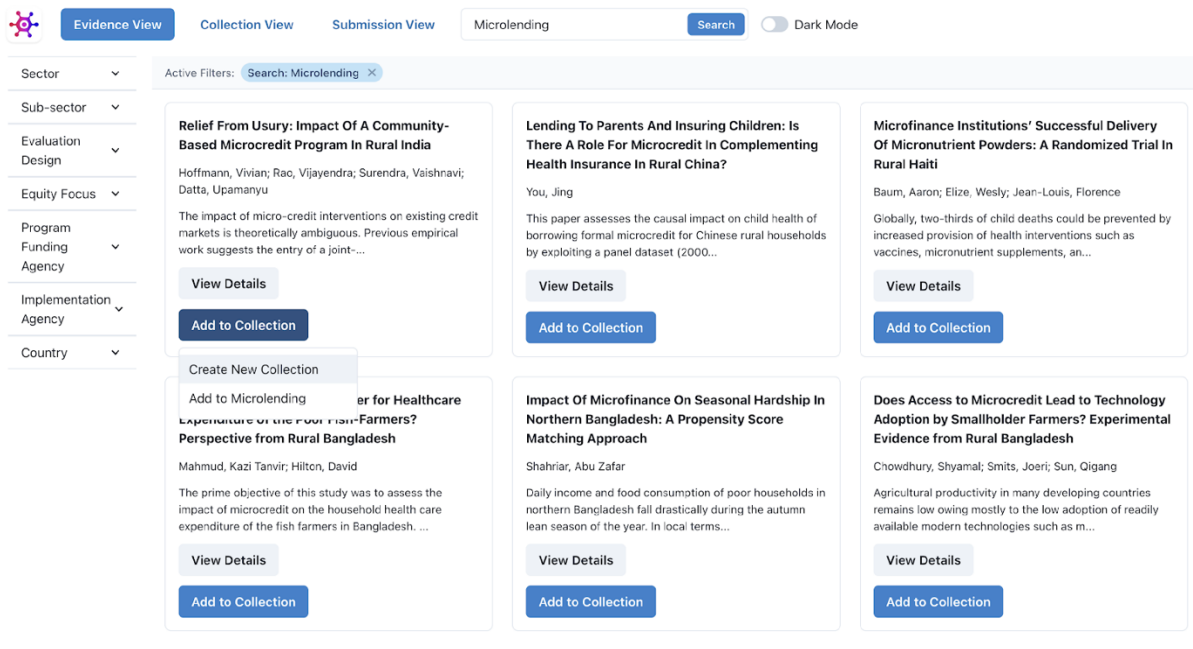
1. **Zoom setup**
 - Confirm audio and video are working.
 - Ask the participant to share their screen – the web browser is sufficient.
2. **Consent**
 - Ensure the consent form is uploaded before proceeding.
3. **InterDev**
 - Go to the InterDev website www.interdev.adaptcentre.ie
 - Ensure the interface is loading correctly. Create a quick collection, go to the collection view and make sure the chat function is working.
 - Verify it responds correctly and is ready to switch to when needed.
 - Log out when you are ready (cmd-k/ctrl-k and then the button in the top right)
 - Check each interface is working correctly (cmd-k/ctrl-k to switch)
 - When the user first opens Interdev, they will be asked to “log in”. **Make sure they use the same pseudonym for this as they did in the demographic survey.**
4. **Final confirmation**
 - Check that both Qualtrics InterDev are working.
 - Be ready to record on Zoom once the tasks begin
 - Be ready to answer participant questions before tasks begin.

Facilitator Instructions – General Introduction

Purpose of this section

Provide the participant with a general introduction of the InterDev system. This is relevant for all conditions of the interface.

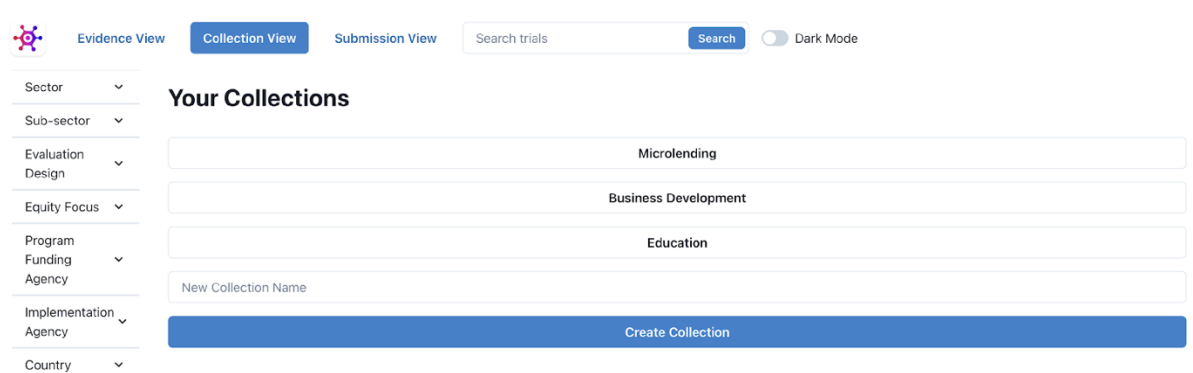
Evidence View



The evidence view is where most individuals will begin their interaction with the system. All trials in the knowledge graph are represented here. Users can search the trials using the searchbox on the top navbar, or do a faceted filter using the navbar on the left side of the screen.

On this screen, users can add trials to a given collection using the “Add to Collection” button. This will also allow them to create a new collection. They can click on the View Details button to learn more about the trial (abstract, authors, country, sector etc). Using the Collection View and Submission View buttons at the top of the screen, users can navigate to different views for the interface. For this evaluation, we are just using the **Evidence View** and **Collection View**. We don’t need to worry about the submission view this time.

Collection View



The collection view is where all of the collections of trials that the user has created live. These are populated from the evidence view. You can make a new collection here and click to view existing collections (eg Microlending, Business Development, Education above).

Collection: Microlending

Turtle [Download Collection](#)

Chat about Microlending

Ask about the trials in this collection... [Send](#)

Relief From Usury: Impact Of A Community-Based Microcredit Program In Rural India
 Hoffmann, Vivian; Rao, Vijayendra; Surendra, Vaishnavi; Datta, Upamanyu
 The impact of micro-credit interventions on existing credit markets is theoretically ambiguous. Previous empirical work suggests the entry of a joint-...

[View Details](#) [Add to Collection](#)

Lending To Parents And Insuring Children: Is There A Role For Microcredit In Complementing Health Insurance In Rural China?
 You, Jing
 This paper assesses the causal impact on child health of borrowing formal microcredit for Chinese rural households by exploiting a panel dataset (2000...

[View Details](#) [Add to Collection](#)

Does Microcredit Really Matter for Healthcare Expenditure of the Poor Fish-Farmers? Perspective from Rural Bangladesh
 Mahmud, Kazi Tanvir; Hilton, David
 The prime objective of this study was to assess the impact of microcredit on the household health care expenditure of the fish farmers in Bangladesh. ...

[View Details](#) [Add to Collection](#)

Clicking on a collection will give the view above. This is where most of the evaluation will be focused on, as the three conditions vary this view. The baseline condition (below) will not have the chat bar. The chat and chat + sources condition (above) will. On this screen, users can explore the trials in the collection and use the same filtering tools as the evidence view. If the chat bar exists, they can use this to learn more and synthesise the data within the collection.

Collection: Microlending

Turtle [Download Collection](#)

Relief From Usury: Impact Of A Community-Based Microcredit Program In Rural India
 Hoffmann, Vivian; Rao, Vijayendra; Surendra, Vaishnavi; Datta, Upamanyu
 The impact of micro-credit interventions on existing credit markets is theoretically ambiguous. Previous empirical work suggests the entry of a joint-...

[View Details](#) [Add to Collection](#)

Lending To Parents And Insuring Children: Is There A Role For Microcredit In Complementing Health Insurance In Rural China?
 You, Jing
 This paper assesses the causal impact on child health of borrowing formal microcredit for Chinese rural households by exploiting a panel dataset (2000...

[View Details](#) [Add to Collection](#)

Does Microcredit Really Matter for Healthcare Expenditure of the Poor Fish-Farmers? Perspective from Rural Bangladesh
 Mahmud, Kazi Tanvir; Hilton, David
 The prime objective of this study was to assess the impact of microcredit on the household health care expenditure of the fish farmers in Bangladesh. ...

[View Details](#) [Add to Collection](#)

Microfinance Institutions' Successful Delivery Of Micronutrient Powders: A Randomized Trial In Rural Haiti
 Baum, Aaron; Elize, Wesly; Jean-Louis, Florence
 Globally, two-thirds of child deaths could be prevented by increased provision of health interventions such as vaccines, micronutrient supplements, an...

[View Details](#) [Add to Collection](#)

Impact Of Microfinance On Seasonal Hardship In Northern Bangladesh: A Propensity Score Matching Approach
 Shahriar, Abu Zafar
 Daily income and food consumption of poor households in northern Bangladesh fail drastically during the autumn lean season of the year. In local terms...

[View Details](#) [Add to Collection](#)

Associational Participation and Network Expansion: Microcredit Self-Help Groups and Poor Women's Social Ties in Rural India
 Davidson, Thomas; Sanyal, Paromita
 Group-based microcredit has become a globally adopted anti-poverty intervention. An important institutional variant of this is the self-help group (SH...

[View Details](#) [Add to Collection](#)

Outside of the tutorial, DO NOT INTERFERE OR HELP the participant if they are stuck. Encourage them to finish the task themselves and if they absolutely cannot figure it out to move on to the next one.

Facilitator Instructions – C0 Condition (Baseline)

Purpose of this section

Introduce participants to the baseline version of the interface. This is as exactly described in the general introduction above – no further introductions necessary.

Facilitator Instructions – C1 Condition (Chat)

Purpose of this section

Introduce participants to the chat interface and ensure they understand how to query and interpret results before attempting the formal tasks.

To introduce the participant to this version of the interface, simply give them an example of using the chat in the collection view, such as that in the image below. Everything else is as per the general introduction.

The screenshot displays a web interface for a microcredit collection. On the left, a sidebar contains filters for Sector, Sub-sector, Evaluation Design, Equity Focus, Program Funding Agency, Implementation Agency, and Country. The main content area is titled 'Collection: Microlending' and shows a 'Turtle' sub-collection with a 'Download Collection' button. Below this is a chat interface titled 'Chat about Microlending'. The chat shows a user asking to design a program incorporating the most salient outcomes from trials, and an assistant responding with three key outcomes: 1. Access to Microcredit, 2. Improved Child Health, and 3. Nutritional Support. Below the chat is a 'Send' button and a text input field. At the bottom, three trial cards are displayed, each with a title, authors, and a brief description. The first card is 'Relief From Usury: Impact Of A Community-Based Microcredit Program In Rural India' by Hoffmann, Vivian; Rao, Vijayendra; Surendra, Vaishnavi; Datta, Upamanyu. The second card is 'Lending To Parents And Insuring Children: Is There A Role For Microcredit In Complementing Health Insurance In Rural China?' by You, Jing. The third card is 'Does Microcredit Really Matter for Healthcare Expenditure of the Poor Fish-Farmers? Perspective from Rural Bangladesh' by Mahmud, Kazi Tanvir; Hilton, David. A 'Mode: Chat Only' button is visible in the bottom right corner of the trial cards section.

Facilitator Instructions – C2 Condition (Chat + KG References)

Purpose of this section

Introduce participants to the chat interface and ensure they understand how to interact with it effectively before attempting the formal evaluation tasks.

There is basically no functional difference in this condition versus the condition with chat only, only that the chat will reference the knowledge graph in its response. So, to introduce the participant to this version of the interface, simply give them an example of using the chat in the collection view again. Everything else is as per the general introduction and chat introduction.

Post-Study Interview and PSSUQ

Directly after completing all of the conditions, the participant will complete a Post Study System Usability Questionnaire (PSSUQ). They will complete this themselves in Qualtrics.

Following this, participants will complete a short post-study interview. This section asks them to reflect on their experience with the **three interface variants**. The questions focus on their preferences, how well each interface supported reasoning, synthesis, and comparison, whether they felt the need to verify claims, and how trustworthy they judged the information provided across interfaces. The questions are below. They will also appear on the Qualtrics survey.

Question Prompt	ID
Of the three interfaces you used (Baseline, Chat, Chat + KG Citations), which did you prefer overall, and why?	Q1
Which interface felt most supportive when you needed to compare or combine findings across multiple evaluations or sources?	Q2
During the tasks, were there particular moments where you felt especially confident in the answer you were producing? Please describe.	Q3
How did the presence or absence of AI support (and citation visibility) influence how much you trusted the information presented?	Q4
Thinking about cognitive effort, which interface felt easiest to work with, and why?	Q5
Is there anything you wanted the system to do that none of the interfaces currently supported?	Q6

Introduce this section by saying:

“You’ve now completed all of the tasks. To finish, we would like you to answer a short set of reflection questions about your experience with the three interfaces. These questions ask about your preferences, how well you felt each interface supported the work, and your impressions of trustworthiness and verification. There are no right or wrong answers — we are interested in your personal perspective. Please take your time and answer as carefully and thoughtfully as you can.”

This section may also be treated as a short semi-structured conversational interview, if appropriate. The facilitator can prompt or follow up on specific parts of the participant’s comments where relevant to trust, verification behaviour, or interface comparison, but should not reinterpret or guide their answers.

“That brings us to the end of the session today. Thank you very much for taking part. Your contributions are extremely valuable to this research. We appreciate the time and effort you’ve put into working through the different systems and tasks. Your responses will help us better understand how people retrieve, evaluate, and synthesise development evidence.”

“Just a reminder that your participation is completely confidential, and if you would like to withdraw your data, you can contact us within and your responses will be removed. If you have any questions about the study, please feel free to reach out to us.”

Compensation Form Email

After completing the evaluation, please send an email with the following form so that we can compensate participants (usually within a month).

https://scsstd.qualtrics.com/jfe/form/SV_7V5CRiPaaD4yrXg

Subject: Compensation Form for Study Participation

Dear [Participant's Name],

Thank you again for taking part in our study. To process your compensation, please complete the short form at the link below:

[Compensation Form](#)

Once submitted, the gift voucher is usually processed within a month.

If you have any questions or run into issues with the form, please don't hesitate to let me know.

Best regards,
[Your Name]

No Show Email

Hi [Participant Name],

Sorry to have missed you today. If you'd still like to take part in our study, you can reschedule at a time that works for you using this link: [Insert Scheduling Link].

We'd be glad to have you participate.

Best regards,
[Your Name]

B. Project Gutenberg Cross Domain Study Protocol

This appendix contains the complete study protocol for the cross-domain evaluation conducted in the digital-library setting. It documents the experimental flow, condition order, task administration procedures, facilitator instructions, and post-condition measures used to compare keyword, semantic, conversational, and conversational-graph interfaces. It relates to Section 5.8 and is included to make clear how the cross-domain study was operationalised as a bounded external-validity probe for the interaction modalities examined in the thesis.

Project Gutenberg (Athena)

Evaluation Protocol V2

Purpose

To evaluate how four different interface designs (Keyword Search, Semantic Search, Chat, Chat+Graph) support users in retrieving and interpreting public domain literature. We measure task performance, workload, subjective impressions and confidence and evaluation behaviour.

Setting

The study will be conducted entirely online using the Qualtrics survey platform, allowing participants to complete it remotely from their own devices. The session is designed to take approximately 40 minutes in total, though participants may take more or less time depending on their pace. All tasks and instructions are self-contained within the survey, and participants will progress independently without direct researcher intervention. This setup ensures a consistent environment across participants while also providing flexibility for them to complete the evaluation in a comfortable setting of their choice.

Central Research Question

Below is the central research question for this thesis. Please keep this in mind throughout the study.

“How does the level of system support, from keyword search through semantic search, LLM chat, and graph-based exploration, affect people’s ability to find, interpret, and synthesize public-domain books, and how does it influence their confidence, trust, and workload?”

Flow Overview

Each participant uses all four interfaces, but the **order is counterbalanced**:

- **Sequence A:** C0 → C1 → C2 → C3
- **Sequence B:** C1 → C2 → C3 → C0
- **Sequence C:** C2 → C3 → C0 → C1
- **Sequence D:** C3 → C0 → C1 → C2

Within each interface block, tasks are randomly drawn from a global pool such that no task is repeated for the same participant. **This is entirely handled by Qualtrics and some custom Javascript, you will not need to do any adjustment for condition assignment.**

1. Introduction & Demographics (6 questions)

Participants begin with a short introduction outlining the purpose of the study, followed by demographic and background questions (e.g., reading habits, familiarity with large language models, awareness of digital libraries).

2. Interface Sequence

Participants are then exposed to four different interface designs. To mitigate order effects, they are randomly assigned to one of four predefined sequences (A–D). This ensures each interface is experienced in a different order across the participant pool.

Each interface sequence consists of:

- **Tutorial Task (1 example task, to be completed by the demonstrator):** A low-stakes practice task to allow participants to become familiar with the interface before the main evaluation.
- **Task Notification:** A reminder that the upcoming tasks are part of the formal evaluation and should be attempted carefully.
- **2 Experimental Tasks:** Participants complete two substantive tasks designed to probe information retrieval, comprehension, and synthesis using the interface. These tasks are drawn without replacement from the global pool, ensuring no participant sees the same task twice across the study.
- **NASA-TLX Workload Assessment (6 sliders) + Serendipity probe:** After completing the two tasks, participants provide self-reported workload ratings using the NASA-TLX dimensions and indicate whether they encountered anything unexpected or interesting during the process.

3. Post-Evaluation Discussion (4 reflective questions)

At the end of the study, participants reflect on their experiences across all four interfaces, discussing preferences, task-interface fit, perceived trustworthiness, and whether they felt the need to verify their answers.

Facilitator Instructions – Study Introduction

The link to the Qualtrics survey is here:

https://scsstd.qualtrics.com/jfe/form/SV_eEzQw4CyCmo3CEm

All study sessions will be conducted over Zoom. Begin by greeting the participant, ensuring their audio connection is clear, and briefly explaining that you will guide them through the study setup before they begin the tasks. You should share the largest monitor available to you, making sure to share your entire screen rather than a single application window. This is essential, as participants will need remote control of your screen in order to interact with the systems.

Before the session, open the Qualtrics survey in a web browser and confirm that it is displayed full screen and ready for the participant to use. The survey introduction page contains the link to the consent form. Explain to the participant that by continuing past the introduction page they are providing implicit consent to take part in the study.

You must also ensure that the Athena system is fully set up in advance. This requires pulling the GitHub project and following the installation and run instructions provided in the repository. Verify that Athena is working correctly in a browser window before the session begins. During the study, participants will interact with Athena directly on your machine, so it is crucial that it is running smoothly and ready to be switched to when needed.

Once you have confirmed that both Qualtrics and Athena are working, share your screen with the participant and hand over remote control. Provide reassurance that they will be guided step by step through each part of the study, and that they can ask clarifying questions at any point. When the participant confirms that they can see and control the screen, direct their attention to the Qualtrics window to begin with the introduction section.

If the user accidentally refreshes or closes the Qualtrics tab, it will automatically reopen where you left off if you return to the survey via the link.

Example Introduction

“Hello, and thank you for joining. Before we begin, I’ll set things up so you can control my screen and complete the study tasks. I’m sharing my entire screen on the largest monitor I have available, which will allow you to interact with both the survey and the experimental systems. In a moment, I’ll give you remote control so you can use my browser directly.”

“On the screen now, you should see the Qualtrics survey. At the top of the introduction page there is a link to the consent form. Please take a moment to read it carefully. Continuing past this page indicates your consent to take part in the study. Participation is voluntary, and you are free to stop at any point.”

“In this study you’ll use four different systems to explore books that are in the public domain. These are works whose copyright has expired or which were never subject to copyright, meaning they are free for anyone to read, share, and use. Each system has a short tutorial, followed by a practice task and then two evaluation tasks. After each system, you’ll answer a short questionnaire about your experience. The entire session should take about 40 minutes. If anything is unclear, feel free to stop and ask me at any time.”

“Before we start, can you see my full screen clearly? And are you able to take remote control of the mouse and keyboard?”

Pre-Study Questionnaire

At the beginning of the Qualtrics survey, participants will be presented with a short questionnaire that collects background information. This includes questions on their reading habits, familiarity with large language models, and awareness of online digital libraries, followed by basic demographic information such as age and gender.

Introduce this section by saying:

“Before we begin the main part of the study, you’ll first answer a few short questions about your reading habits, your familiarity with language models and digital libraries, and some demographic details. These are standard background questions and will only

take a few minutes to complete. Please answer them as honestly as you can. Once you have finished, the survey will guide you directly into the first interface tutorial.”

Ensure that the participant understands that these questions are for research context only, that their responses are confidential, and that they may skip any question they do not wish to answer (e.g., “Prefer not to say” options).

NB Once the Pre-Study Questionnaire is complete please advise the participant that you will be now turning on the recording on Zoom and ensure that they are aware and comfortable with this. Then don’t forget to turn it on!

Facilitator Pre-Session Checklist

1. Zoom setup

- Confirm audio and video are working.
- Share your **entire screen** (not just a window).
- Enable remote control for the participant.
- Use the largest monitor available.

2. Qualtrics survey

- Open the Qualtrics survey in a browser.
- Ensure the introduction page is visible and full screen.
- Verify that the consent form link is present.

3. Athena system

- Pull the [GitHub project](#) to the local machine if not done already.
- Follow the setup instructions from the repository.
- Launch Athena in a browser window.
- Verify it responds correctly and is ready to switch to when needed.
- Check each interface is working correctly (cmd-k/ctrl-k to switch)

4. Final confirmation

- Check that both Qualtrics and Athena windows are open and working.
- Make sure screen sharing shows the correct monitor.
- Be ready to record on Zoom once the tasks begin
- Be ready to answer participant questions before tasks begin.

Facilitator Instructions – C0 Condition (Project Gutenberg Keyword Search)

Purpose of this section

Introduce participants to the **Project Gutenberg keyword search interface** and ensure they can carry out simple searches before attempting the formal tasks.

Step 1: Introduction

Say to the participant:

“In this part of the study, you’ll be using the Project Gutenberg digital library through its keyword search function. This interface allows you to look up works in the collection by entering words or phrases. I’ll walk you through the main features, and then you’ll have the chance to try a short practice task before moving on to the real evaluation tasks.”

Step 2: Navigating the Interface

Explain the main page layout:

- **Search bar:** Located at the top; type in keywords (e.g., author names, titles, or themes).
- **Search results list:** Shows matches with links to book information pages.
- **Book information page:** Provides metadata (title, author, publication date, etc.) and download links.
- **Download options:** Typically include plain text (.txt) and other formats such as EPUB or Kindle.

Step 3: Search Tips

Advise the participant that:

- Simple keyword searches (e.g., “*Shakespeare*” or “*detective*”) usually work best.
- They can be refined by author (e.g., “*Mark Twain*”) or by topic keywords (e.g., “*friendship*”, “*travel*”).
- More advanced options such as *Category search* and *Advanced search* are available.

Step 4: Example Demonstration

- Facilitator demonstrates:
 1. Enter a known title, e.g., “The Power of Four.”
 2. Show how the result appears and click through to the book page.
 3. Point out the download links (plain text, EPUB, etc.).
- Say:

“This is the workflow you’ll use in your tasks: search, review results, navigate to the book, and extract the requested information.”

Step 5: Practice Task

- Give the participant a low-stakes practice task. This will be displayed on screen at this point on the Qualtrics survey, eg:
“Try searching for a novel that deals centrally with poverty or social deprivation. When you find one, record the title. You don’t need to spend long on this, it’s just to get familiar with the system.”
- Allow them to work; answer clarifying questions if they are stuck.
- Emphasize: *“This is only practice, you can move on whenever you feel ready.”*
- Let the user know that the confidence scoring is rated out of 100 with 0 being not confident in their answer whatsoever and 100 being absolutely confident in their answer.

Step 6: Transition to Evaluation

Say:

“Now that you’ve seen how the interface works, you’ll move on to two formal tasks. Please complete these to the best of your ability. If you don’t feel you can finish one, you can continue to the next. Afterward, you’ll answer a short workload and reflection questionnaire.”

Outside of the tutorial, DO NOT INTERFERE OR HELP the participant if they are stuck. Encourage them to finish the task themselves and if they absolutely cannot figure it out to move on to the next one.

Facilitator Instructions – C1 Condition (Semantic Search)

Purpose of this section

Introduce participants to the semantic search interface and ensure they understand how to query and interpret results before attempting the formal tasks.

Step 1: Introduction

Say to the participant:

“In this part of the study, you’ll be using a semantic search interface to explore Project Gutenberg. Semantic search lets you enter a phrase or question in natural language, and it will retrieve books that are conceptually related. I’ll walk you through how to use it, and then you’ll have a chance to try a short practice task before starting the evaluation tasks.”

Step 2: Navigating the Interface

Explain the main features:

- **Search box:** Enter queries in full sentences or descriptive phrases.
- **Results list:** Shows books ranked by semantic relevance to the query.
- **Book information page:** Includes title, author, description, and download links.

- **Semantic Richness:** Results may include related works even if the query terms are not explicitly in the title.

Step 3: Search Tips

Advise the participant that:

- Queries can be descriptive (e.g., “*a story about friendship during wartime*”).
- Short queries also work (like “*ghost stories*” or “*Charles Dickens*”), but the system often works best when phrased naturally.
- If results don’t look right, they can rephrase the query using different words or phrasing.

Step 4: Example Demonstration

Facilitator demonstrates a search and talks it through:

1. Enter a natural language query, e.g., “*A novel about a scientist who creates life.*”
2. Show how results differ from keyword search — related works appear even if the words don’t exactly match the query.
3. Click through to a book page and point out metadata and download options.

Say:

“This is the workflow you’ll use in your tasks: write a query in your own words, review the results, select a relevant work, and extract the information needed.”

Step 5: Practice Task

- Present the tutorial task (displayed in Qualtrics):
“*Identify a novel that is commonly described as one of the first feminist novels in English and is available on Project Gutenberg. Provide the title and briefly explain why it supports being called a feminist text.*”
- Encourage them to try queries in their own words.
- Emphasize: “*This is just practice. You don’t need to be perfect, continue when you feel ready.*” Encourage them to explore the system openly before continuing.
- Let the user know that the confidence scoring is rated out of 100 with 0 being not confident in their answer whatsoever and 100 being absolutely confident in their answer.
-

Step 6: Transition to Evaluation

Say:

“Now that you’ve had practice with semantic search, you’ll begin two evaluation tasks. Please complete them to the best of your ability. If you don’t feel you can finish one, you can continue to the next. Afterward, you’ll answer a short workload and reflection questionnaire.”

Outside of the tutorial, DO NOT INTERFERE OR HELP the participant if they are stuck. Encourage them to finish the task themselves and if they absolutely cannot figure it out to move on to the next one.

Facilitator Instructions – C2 Condition (Chat Interface)

Purpose of this section

Introduce participants to the chat interface and ensure they understand how to interact with it effectively before attempting the formal evaluation tasks.

Step 1: Introduction

Say to the participant:

“In this part of the study, you’ll be using a chat-based system to explore works of literature in the public domain. The system is designed to answer questions, suggest works, and help you locate information by responding to your typed queries. I’ll show you the main features, and then you’ll try a short practice task before starting the evaluation tasks.”

Step 2: Navigating the Interface

Explain the main features:

- **Chat input box:** Type questions or prompts in natural language. The users can explicitly request books eg. *“I’m looking for books about fishing”* or *“Get me books featuring characters overcoming poverty in the 19th century”* and then ask follow up questions about the books that were fetched eg. *“Which of these books would pair well for an introductory course on Victorian literature?”* or *“Can you rank these books by how easy they are to read?”* If the system misinterprets the users request for books, they can use the “Add Books” button to force the system to search for books.
- **Conversation thread:** The system replies in text, showing search results, book titles, or explanatory information.
- **Interaction style:** The conversation builds over turns; participants can ask follow-up questions to refine or clarify results.
- **Book access:** When the system suggests a title, participants can follow the link or reference to access metadata and downloads.
- **Get Books:** The system should automatically pick up when a participant is looking for books vs when they are just asking a question. If for some reason it is not, clicking on this option will force the system to search for books.

Step 3: Query Tips

Advise the participant that:

- They can phrase queries naturally, like asking a question or making a request.

- If the first response doesn't fully answer their question, they can continue the conversation by rephrasing or asking follow-up questions.
- The chat is flexible: it can provide titles, summaries, or thematic suggestions depending on how the question is asked.

Step 4: Example Demonstration

Facilitator demonstration (screen-share or walk-through):

1. Type a general request such as *"Recommend a novel about seafaring adventures."*
2. Show how the chat produces a reply, possibly including relevant works.
3. If needed, follow up with *"Can you give me more details about the author?"* to show how the conversation can continue.

Say:

"This is the workflow you'll use in your tasks: type your request, review the response, and ask follow-up questions until you find the information you need."

Step 5: Practice Task

- Present the tutorial task shown in Qualtrics at this point, eg:
"Find a work that was originally written in French and for which an English translation is available. Provide the title and the author's name."
- Let the participant experiment with conversation turns.
- Emphasize: *"This is just practice. You can move on whenever you feel ready."*
- Let the user know that the confidence scoring is rated out of 100 with 0 being not confident in their answer whatsoever and 100 being absolutely confident in their answer.

Step 6: Transition to Evaluation

Say:

"Now that you've practiced using the chat system, you'll move on to two formal tasks. Please complete these to the best of your ability. If you don't feel you can finish one, you may continue to the next. Afterward, you'll answer a short workload and reflection questionnaire."

Outside of the tutorial, DO NOT INTERFERE OR HELP the participant if they are stuck. Encourage them to finish the task themselves and if they absolutely cannot figure it out to move on to the next one.

Facilitator Instructions – C3 Condition (Chat + Graph Interface)

Purpose of this section

Introduce participants to the chat and graph-based system ("Athena") and ensure they

understand how to combine conversational queries with the graph visualization before attempting the formal evaluation tasks.

Step 1: Introduction

Say to the participant:

“In this part of the study, you’ll be using a system that combines a chat interface with a visual graph. The chat allows you to ask questions and make requests in natural language, while the graph shows relationships between books, authors, and themes. I’ll show you how the two work together, and then you’ll have a chance to try a short practice task before moving on to the evaluation tasks.”

Step 2: Navigating the Interface

Explain the main features:

- **Chat input box:** Type questions or prompts in natural language. The users can explicitly request books eg. *“I’m looking for books about fishing”* or *“Get me books featuring characters overcoming poverty in the 19th century”* and then ask follow up questions about the books that were fetched eg. *“Which of these books would pair well for an introductory course on Victorian literature?”* or *“Can you rank these books by how easy they are to read?”* If the system misinterprets the users request for books, they can use the “Add Books” button to force the system to search for books.
- **Graph area:** Displays entities such as books, authors, or concepts, and how they are connected.
- **Interaction:** Clicking on graph nodes can reveal more details or focus the view.
- **Book details panel:** Provides metadata (title, author, summary) and download links.
- **Graph controls:** Options such as deleting nodes or simplifying the view if it becomes cluttered.
- **Summarise:** You can summarise the books by selecting them in graph view (shift+click) and then clicking on the summarise button.
- **Context Chat:** You can chat about books with context by selecting them in graph view (shift+click) and then chatting in the box to the left.
- **Get Books:** The system should automatically pick up when a participant is looking for books vs when they are just asking a question. If for some reason it is not, clicking on this option will force the system to search for books.

Step 3: Usage Tips

Advise the participant that:

- They can start by asking broad or descriptive questions, and the graph will expand to show related works.
- Following links in the graph or clicking on nodes can help them explore connections that aren’t immediately obvious.

- They can mix chat and graph interactions for example, asking a new question after clicking a node to refine the context.
- If the graph becomes crowded, they can delete nodes to keep it manageable.

Step 4: Example Demonstration

Facilitator demonstration:

1. Type a general query such as *“Show me novels about travel.”*
2. Point out how the graph updates with relevant works and how nodes are connected.
3. Click on a node to reveal details (title, author, and download links).
4. Show how to remove a node from the graph to tidy the workspace.
5. Show how to create a summary from selected books on the graph
6. Show how to ask questions about selected books on the graph

Say:

“This is the workflow you’ll use: ask a question, review both the chat reply and the graph expansion, and interact with nodes to explore further or find the information you need.”

Step 5: Practice Task

- Present the tutorial task (displayed in Qualtrics) eg:
“Write a concise explanation (80–120 words) of what ‘Gothic literature’ is. Base your explanation on at least two works that are commonly considered Gothic.”
- Allow the participant to try both the chat and graph interactions while searching for supporting texts. Encourage exploration of the system.
- Emphasize: *“This is just practice, continue when you feel ready.”*
- Let the user know that the confidence scoring is rated out of 100 with 0 being not confident in their answer whatsoever and 100 being absolutely confident in their answer.

Step 6: Transition to Evaluation

Say:

“Now that you’ve practiced using the chat-and-graph system, you’ll begin two evaluation tasks. Please complete them to the best of your ability. If you don’t feel you can finish one, you may continue to the next. Afterward, you’ll answer a short workload and reflection questionnaire.”

Outside of the tutorial, DO NOT INTERFERE OR HELP the participant if they are stuck. Encourage them to finish the task themselves and if they absolutely cannot figure it out to move on to the next one.

Post-Study Questionnaire

At the end of the Qualtrics survey, participants will complete a short post-study questionnaire. This section asks them to reflect on their experience with the four interfaces. The questions cover which interface they preferred and why, which tasks they felt suited each interface, whether they felt the need to verify their answers, and how trustworthy they judged the responses from each system.

Introduce this section by saying:

“You’ve now completed all of the tasks. To finish, we’d like you to answer a short set of reflection questions about your experience. These questions ask about your preferences, how well you felt each interface supported different kinds of tasks, and your impressions of trustworthiness. There are no right or wrong answers here. We’re interested in your personal perspective. Please take your time and answer as thoughtfully as you can.”

You can treat this section as a semi-structured interview and facilitate discussion in topics that concerning the interface that the user is interested in. Once they submit this section, their participation in the study is complete.

“That brings us to the end of today’s session. Thank you very much for taking part. Your contributions are really valuable to our study. We appreciate the time and effort you’ve put into working through the different systems and tasks. Your responses will help us better understand how people search for and interact with digital library collections.”

“Just a reminder that your participation is completely confidential, and if you would like to withdraw your data, you can contact us within and your responses will be removed. If you have any questions about the study, please feel free to reach out to us.”

Compensation Form

After completing the evaluation, please send an email with the following form so that we can compensate participants (usually within a month).

https://scsstcd.qualtrics.com/jfe/form/SV_9Bv7IfhYrvWVbYq

Subject: Compensation Form for Study Participation

Dear [Participant’s Name],

Thank you again for taking part in our study. To process your compensation, please complete the short form at the link below:

[Compensation Form](#)

Once submitted, the gift voucher is usually processed within a month.

If you have any questions or run into issues with the form, please don’t hesitate to let me know.

Best regards,
[Your Name]

No Show Form

Hi [Participant Name],

Sorry to have missed you today. If you'd still like to take part in our study, you can reschedule at a time that works for you using this link: [Insert Scheduling Link].

We'd be glad to have you participate.

Best regards,
[Your Name]

When the study is completed and you have received the Zoom recording link, please update the following sheet:

[Recording links](#)

C. End-to-End Prompt Example for Project Gutenberg Retriever

This appendix provides a worked example of the retrieval and prompting pipeline used in the Project Gutenberg cross-domain study. It shows how a user query is transformed into a structured intent, how candidate items are selected, how grounded explanations are generated, and how graph outputs are constructed. It relates to Section 5.8.2 by giving concrete implementation detail for the retrieval-augmented interaction pipeline underlying the semantic, conversational, and conversational-graph conditions.

User Request

A client POSTs to /query_books_graph with {"query": "Looking for three witty English Victorian novels by women authors who critique society", "n": 3, "mode": "graph"} plus the Bearer API key. The handler validates the payload, loads cached subjects, and starts intent parsing before contacting Gutendex.

Structured Intent

The parse_structured_intent helper prompts gpt-4o-mini with the CSV of supported subjects and extracts normalized languages, topics, search_terms, and optional author-year bounds. For this query the model could return:

```
{
  "languages": ["en"],
  "topics": [
    "English fiction -- 19th century",
    "Women authors",
    "Social problems -- Fiction"
```

```

],
"search_terms": [
  "jane austen",
  "elizabeth gaskell",
  "victorian romance"
],
"author_year_start": 1800,
"author_year_end": 1899
}

```

Using those signals, the API builds deterministic Gutendex requests per topic/language/year.

Candidate Summary Sent to GPT

After combining popularity-ranked and random topic coverage lists, each remaining book compresses to "ID | Title | Author | sig=...", where sig is the normalized title signature. For this scenario the snippet looks like:

```

1342 | Pride and Prejudice | Jane Austen | sig=prideandprejudice
1465 | Cranford | Elizabeth Cleghorn Gaskell | sig=cranford
969 | The Tenant of Wildfell Hall | Anne Bronte | sig=thetenantoftwildfellhall
768 | Lady Susan | Jane Austen | sig=ladysusan
158 | Emma | Jane Austen | sig=emma

```

Those lines are injected into the selector prompt:

System:

You are a precise API that selects book IDs. Given a list of candidates in the form 'ID | Title | Author | sig=...' and a user query, return ONLY {"selected_book_ids": [IDs]} with exactly 3 unique integers. Rules: (1) prefer exact sig matches; (2) next prefer sig prefixes; (3) otherwise choose best semantic matches; (4) JSON only.

User:

User query: Looking for three witty English Victorian novels by women authors who critique society
 query_sig=lookingforthreewittyenglishvictoriannovelsbywomenauthorswhocritiquesociety

Candidates:

```

1342 | Pride and Prejudice | Jane Austen | sig=prideandprejudice
1465 | Cranford | Elizabeth Cleghorn Gaskell | sig=cranford
969 | The Tenant of Wildfell Hall | Anne Bronte | sig=thetenantoftwildfellhall
768 | Lady Susan | Jane Austen | sig=ladysusan
158 | Emma | Jane Austen | sig=emma

```

Select exactly 3 IDs. JSON only.

GPT-4o-mini replies with something like `{"selected_book_ids":[1342,1465,969]}`, which the backend normalizes, deduplicates, and backfills if underfilled (gutget/app.py:1099-1135).

User-Facing Explanation Prompt

Only the chosen titles feed the explanation prompt so the response is grounded in actual selections. The call to gpt-4o uses:

System:

You are a concise, friendly librarian. One short paragraph (100–140 words). No IDs.

User:

User query: Looking for three witty English Victorian novels by women authors who critique society

You, the librarian, have just chosen these titles from public domain literature: ['Pride and Prejudice', 'Cranford', 'The Tenant of Wildfell Hall']

Explain why these fit. Mention that the books have been added to the graph on the right for download or summaries, and that the user can click to highlight multiple books.

RULES: ONLY if none fit exactly, explain why and cite alternatives.

A representative completion:

"I've added Pride and Prejudice, Cranford, and The Tenant of Wildfell Hall to the graph on the right because each is an English, woman-authored Victorian novel that skewers social constraints with wit. Austen's classic exposes the marriage market as comedic theater, Gaskell's Cranford gently satirizes provincial manners through its resilient women, and Brontë's Tenant tackles gendered double standards head-on. Click any node to pull summaries or downloads, and highlight multiple books to compare how each author critiques class and gender"

Graph Construction

The selected metadata are transformed into nodes with `gutindex_id`, `cover`, `languages`, and `subjects`. It prompts GPT for JSON edges with natural-language reasons and hub-and-spoke grouping so every edge has readable provenance. For this example the graph payload might be:

```
{
  "nodes": [
    {"id": "a", "data": {"gutindex_id": 1342, "title": "Pride and Prejudice", "author": "Jane Austen", "subjects": ["Young women -- Fiction", "Courtship -- Fiction"], "book_download": "https://.../1342-0.txt"}}],
```

```

    {"id":"b","data":{"gutindex_id":1465,"title":"Cranford","author":"Elizabeth Cleghorn
Gaskell","subjects":["England -- Social life and customs -- 19th
century"],"book_download":"https://.../1465-0.txt"}},
    {"id":"c","data":{"gutindex_id":969,"title":"The Tenant of Wildfell
Hall","author":"Anne Brontë","subjects":["Women -- Social conditions","Didactic
fiction"],"book_download":"https://.../969-0.txt"}}
  ],
  "edges": [
    {"id":"rel0","source":"a","target":"b","data":{"reason":"These novels are English
Victorian satires written by women examining class and manners."}},
    {"id":"rel1","source":"a","target":"c","data":{"reason":"Both works challenge gender
expectations through witty heroines transforming their communities."}}
  ]
}

```