Cross-Site Personalisation

A thesis submitted to the

University of Dublin, Trinity College

For the degree of

Doctor of Philosophy

Kevin Koidl

Knowledge and Data Engineering Group,

Department of Computer Science,

Trinity College,

Dublin

2013

Declaration

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

_____

Kevin Koidl

April 2013

Permission to Lend or Copy

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

_____

Kevin Koidl

April 2013

Acknowledgements

"It is not the destination that makes the journey worthwhile."

Proverb

This journey was full of adventures beyond my wildest expectations and it was the support and guidance of the people around me that allowed me to overcome the many obstacles along the way.

Firstly, I would like to thank my supervisor, Prof. Owen Conlan, whose knowledge and encouragement has guided me through this research. Special thanks are also reserved for Prof. Vincent Wade, my co-supervisor, whose support and insightful advice was invaluable.

I would like to thank my parents for their unconditional and unfailing love, encouragement, belief and guidance throughout my life. You are my role models and have set an example that I will always aspire to live up to. My brother Roman for always being there when I needed him. My wife Niuscha, whose love, kindness and encouragement gave me balance when things got hard.

I would also like to thank my colleagues and friends in the Knowledge and Data Engineering Group for helping me along the way. My special thanks goes to John McAuley, Brian Gallagher, Dr. Cormac Hampson, Luca Lungo, Dr. Alex O'Connor, Prof. Seamus Lawless, Dr. Dominic Jones and Hilary McDonald.

I also would like to thank my wider circle of friends assisting me with the write up of this thesis, especially Dr. Emmett P. Tracey and Dr. Bahman Honari.

Finally, I would like to express my gratitude to Science Foundation Ireland for funding the research detailed in this thesis through the Centre of Next Generation Localisation (CNGL).

I sincerely thank you all

Es gibt noch eine andere Welt zu entdecken – und

mehr als eine!

Auf die Schiffe, ihr Philosophen!"

"There is still another world to discover –

and more than one!

Aboard ship, ye philosophers!"

*(*Friedrich Nietzsche, *Die fröhliche Wissenschaft – The Joyful Wisdom*)

Abstract

Web users are continuously confronted with vast amounts of information. This phenomenon is known as the *information explosion* and can lead to disorientation and decreased productivity as users attempt to navigate through information to discover what they need. Web Personalisation techniques, such as Adaptive Hypermedia, have been employed to overcome this issue. However, these techniques are typically focussed on closed collections of content, housed on a single website. This is at odds with the typical web browsing paradigm, where users navigate across multiple sites to identify relevant information.

This thesis introduces the Cross-Site Personalisation (CSP) approach to Web Personalisation. This approach seamlessly personalises the support offered to each individual user as they browse across multiple websites. The approach models the user's interactions and then tailors the information on each independent website to support their needs. The Cross-Site Personalisation approach is realised as a third-party personalisation service that interfaces with multiple websites via standard extension modules. The approach is non-intrusive and does not negatively impact on the user's web browsing experience. It is also sensitive to the user's privacy concerns by not divulging information about a user to the websites being personalised. Rather, this approach recommends how each website may adapt their information and navigation structures to meet the user's needs. To evaluate the appropriateness of this approach two authentic use cases have been scoped, designed and developed. The first use case focuses on CSP within closed domains, such as those typically seen in enterprise website federations. The second use case addresses open domains, where the user is navigating across websites on the open web.

This thesis details the State of the Art that underpins the Cross-Site Personalisation approach. It provides a detailed design and corresponding implementations that realise the two use cases. Furthermore, this thesis presents a thorough evaluation of the applicability and appropriateness of the approach in these authentic settings.

Table of Contents

Declaration ii

Permission to Lend or Copy iii

Acknowledgements iv

Abstract vi

1. Introduction 1

1.1. Motivation 1

1.2. Research Question 4

1.3. Research Objective, Goals and Challenges 5

1.4. Methodology 6

1.5. Contribution 6

1.6. Scope 8

1.7. Thesis Overview 8

2. State of The Art Review 10

2.1. Introduction 10

2.2. Information Overload and the Limitations of the Current Search Paradigm 13

2.3. Gap in the State of the Art 15

2.4. Web User Dimension 17

2.4.1. Introduction 17

2.4.2. Understanding and Assisting the Web Users in Addressing Information Needs 17

2.4.3. Modelling Information Needs 24

2.4.4. User Identification, Online Privacy and Trust 29

2.4.5. Summary 30

2.5. Web Content Dimension 33

2.5.1. Introduction 33

2.5.2. Extracting Meaning from Pre-Structured Web Content 33

2.5.3. Extracting Meaning from Unstructured Web Content 35

2.5.4. Closed and Open Corpus 37

2.5.5. Summary 38

2.6. Web Application Dimension 39

2.6.1. Introduction 39

2.6.2. Non-Intrusive Web Personalisation Techniques 40

2.6.2.1. Information Retrieval 41
2.6.2.2. Recommender Systems 42
2.6.2.3. Rule-based Systems 44
2.6.3. Non-Invasive Integration of Web Personalisation Techniques 45
2.6.3.1. Non-Invasive integration of Web Personalisation at design-time 46
2.6.3.2. Non-Invasive integration of Web Personalisation at run-time 47
2.6.4. Interoperability of Web Personalisation Approaches 49
2.6.4.1. Adaptive Hypermedia Systems 50
2.6.4.2. Scout approaches 53
2.6.4.3. Intermediary Approaches (Portal approaches) 54
2.6.4.4. Browser Extensions/Plug-ins 55
2.6.5. Summary 56
2.7. Conclusions 59
3. Design 60
3.1. Influences from the State of the Art 60
3.1.1. Web User Dimension Requirements 60
3.1.2. Web Content Dimension Requirements 62
3.1.3. Web Application Dimension Requirements 62
3.1.4. Summary 63
3.2. High-Level Design 65
3.2.1. High-Level Design Considerations 65
3.2.2. CSP Architecture Overview and Component Responsibilities 67
3.2.3. Summary 71
3.3. Cross-Site Personalisation in Closed Domains 74
3.3.1. Characteristics of a Closed Domains in the Context of CSP 74
3.3.2. CSP Design Considerations for Use cases in Closed Domains 75
3.3.3. CSP Architecture Components and Responsibilities in Closed Domains 76
3.3.4. Summary 79
3.4. Cross-Site Personalisation in Open Domains 84
3.4.1. Characteristics of a Closed Domains in the Context of CSP 84
3.4.2. Design Considerations for CSP Use cases in Open Domains 85
3.4.3. CSP Architecture Components and Capabilities 86
3.4.4. Summary 88

3.5. Conclusions 94
4. Implementation 96
4.1. Prototype Implementation for Cross-Site Personalisation in Closed Domains 96
4.1.1. Closed Domain Characteristics 97
4.1.2. Use case Example and Pre-Requisites 97
4.1.3. Content Considerations 99
4.1.4. Technological Architecture and Components 102
4.1.5. Process Overview 110
4.1.6. Summary 111
4.2. Prototype Implementation for Cross-Site Personalisation in Open Domains 118
4.2.1. Open Domain Characteristics 118
4.2.2. Use case Example and Pre-Requisites 118
4.2.3. Content Considerations 120
4.2.4. Technological Architecture and Components 121
4.2.5. Process Overview 130
4.2.6. Summary 131
4.3. Conclusion 138
5. Evaluation 140
5.1. Chapter Overview 140
5.2. Initial User Survey 143
5.2.1. Survey Objectives and Setup 143
5.2.2. Findings 143
5.2.3. Discussion 149
5.3. Cross-Site Personalisation in Closed Domains 151
5.3.1. Experiment Objectives 151
5.3.2. Experimental Setup 151
5.3.3. Baseline Comparison 153
5.3.4. Results 154
5.3.5. Discussion 163
5.4. Two-Phased User Study: CSP-based Visual Assistance and Relevance 165
5.4.1. Phase 1: Investigation of Link-based Visual Assistance 165
5.4.2. Phase 2: Behavioural Tracking and Recommendation Creation 168
5.4.3. Discussion 168

5.5. Cross-Site Personalisation in Open Domains 170
5.5.1. Experiment Objectives 170
5.5.2. Experimental Setup 170
5.5.3. Baseline Comparison 172
5.5.4. Results User Performance 173
5.5.5. Results Self-Reported Feedback 178
5.5.6. Further Findings 183
5.5.7. Discussion 185
5.6. System Related Evaluation Considerations 188
5.7. Concluding Discussion 191
6. Conclusion 196
6.1. Achievements 196
6.2. Contribution to the State of the Art 204
6.3. Future Work 206
6.3.1. User Profile Management and Social Media 206
6.3.2. Term Identification 207
6.3.3. Information Need Detection 208
6.3.4. Flexible Intrusiveness 208
6.3.5. Mobile and Localisation 209
6.3.6. Web Narratives 210
Bibliography 212
Appendices 226
Appendix I – State of the Art Survey of Web-based Content Management Systems 226
Introduction 226
Survey Outline 227
Comparison of Web Content Management Systems 230
MediaWiki 237
References of Study 241
Appendix II – Online Survey to investigate the browsing experience of users in an across different Web-based Content Management Systems 243
Section I: General questions about browser usage 243
Section II: General Questions about Web-based Content Management System usage 243

Section III: Questions related to the usage of different WCMS for the same interest/goal/intent 246
Section IV: Questions about personalisation in WCMS 247
Appendix III – First Experiment (Closed Domains) 250
A.1. Tasks 250
A.2. Task performance metrics (through logging) 253
A.3. Pre-test Questionnaire 254
A.4. Usability Questionnaires 256
A.5 Additional Questions 261
Selected data outputs 263
Appendix IV – Two-Phased Design Study: Relevance, Usability and Visual Assistance 265
Phase 1 Questionnaires: Investigation of Visual Assistance Types 265
Phase 1 Findings: Investigation of Visual Assistance Types 270
Phase 2 Questionnaire: Behavioural Tracking and Recommendation Creation 274
Phase 2 Findings: Behavioural Tracking and Recommendation Creation 282
Appendix V – Second Experiment (Open Domains) 288
A.1. Tasks 288
A.2. Task performance metrics (through logging) 293
A.3. Pre-test Questionnaire 295
A.4. Post Questionnaires 296
A.5. Concluding questions and comments to overall personalisation approach 298
Findings User Performance 299
Findings Self-Reported Feedback 307
Integration Overview of CSP in Open Domains Part 1 (Figure is split into two parts) 313
Integration Overview of CSP in Open Domains Part 2 (Figure is split into two parts) 314

List of Figures

Figure 1 Web Personalisation Dimensions....12
Figure 2 Taxonomy of Adaptive Hypermedia assistance mechanisms (Brusilovsky, 2001, 1996) ...22
Figure 3 Overarching CSP application areas to address....40
Figure 4 High-Level Service View....66
Figure 5 High-Level CSP Architecture....68
Figure 6 High-level Example Use case of CSP in a Closed Domain Environment....98
Figure 7 Concept Ontology (excerpt) based on DocBook extraction....101
Figure 8 Technical CSP Architecture for Closed Domain....103
Figure 9 CSP interaction example in relation to search....106
Figure 10 Example of Personalised Link Annotations in Drupal....108
Figure 11 CSP Closed Domain Website Integration Process....109
Figure 12 Process Overview Closed Domain Use case....111
Figure 13 High-level Example CSP Use case in a Open Domain....119
Figure 14 Technical CSP Architecture for Open Domains....122
Figure 15 Process Overview for Open Domain Use cases....130
Figure 16 Evaluation Overview....141
Figure 17 Usages of WCMS....144
Figure 18 Content Type Preferences....144
Figure 19 The use of the same WCMS implementation for the same interest/goal/intent....145
Figure 20 Information Overload Related Phenomena....145
Figure 21 Usefulness of personalisation in and across WCMS....146
Figure 22 Rating of Assistance Type....147
Figure 23 Usages of Browser Plug-Ins....147
Figure 24 Creation of User Profile based on Browsing Behaviour....148
Figure 25 Storing of User Profile in WCMS....148
Figure 26 Allowing a third-party Service to Access User Model....148
Figure 27 Allowing a third-party Service to use a User Model across different WCMS....149
Figure 28 Cross-Site Personalisation indicated as Link Annotation....154
Figure 29 Task Completion Time Closed domain use case Evaluation....155
Figure 30 Overall Mean Task Completion Time Closed Domain Use case Evaluation....156
Figure 31 Overall Clicks per Task and Website Closed Domain Use case Evaluation....157
Figure 32 Total Clicks per System Setting Closed Domain Use case Evaluation....157
Figure 33 Box Chart Overall Clicks Closed Domain Use case Evaluation....158
Figure 34 Self-Reported Findings Closed Domain Use case Evaluation Related to Relevancy of Adaptive Hints....159
Figure 35 Self-Reported Findings Closed Domain Use case Evaluation: Adaptive Hints In Relation To Adaptive Hint Usefulness....160
Figure 36 Self-Reported Findings Closed Domain Use case Evaluation: Awareness of Adaptive Hints....160
Figure 37 Self-Reported Findings Closed Domain Use case Evaluation: Feeling of being lost....161
Figure 38 Link Highlighting....166
Figure 39 Link Annotation with a star symbol....166
Figure 40 Link Annotation with a circular symbol....166
Figure 41 Small Dotted Bar....166

Figure 42 Example of CSP Based Annotation on a website 173
Figure 43 Mean Clicks Each Websites Open Domain Use case Evaluation 174
Figure 44 Mean Overall Clicks Open Domain Use case Evaluation 175
Figure 45 Mean Clicks Killarney and Dingle Open Domain Use case Evaluation 175
Figure 46 Overall Mean Times Open Domain Use case Evaluation 176
Figure 47 Mean Time Killarney and Dingle Open Domain Use case Evaluation 177
Figure 48 Box Plot Task Completion Time Open Domain Use case Evaluation 177
Figure 49 Self-Reported Frustration Open Domain Use case Evaluation 179
Figure 50 Self-Reported Motivation Open Domain Use case Evaluation 179
Figure 51 Self-Reported Task Relevance Open Domain Use case Evaluation 180
Figure 52 Self-Reported Usefulness of System Guidance Open Domain Use case Evaluation 181
Figure 53 Self-Reported Difficulty in Browsing Open Domain Use case Evaluation 181
Figure 54 Self-Reported Performance Open Domain Use case Evaluation 182
Figure 55 Self-Reported Privacy Concerns Open Domain Use case Evaluation 183
Figure 56 Overall Average Time by Category Open Domain Use case Evaluation 184
Figure 57 Average Time per Category Killarney and Dingle Open Domain Use case Evaluation 184
Figure 58 Drupals Layer Architecture (Drupal, 2009) 233
Figure 59 Illustrating Drupals Technology Stack (vanDyk, 2008) 234
Figure 60 An Overview of Drupals core (vanDyk, 2008) 234
Figure 61 Drupals Semantic Web Architecture (http://semanticsearch.org/) 236
Figure 62 Sematic MediaWiki (Volkel, 2006) 240
Figure 63 Mean Total Links Visited All Websites 299
Figure 64 Box chart Total Links Visited 300
Figure 65 Box chart Total Links Vi sited Killarney and Dingle 301
Figure 66 Line Chart Average Relevant Links Visited 301
Figure 67 Line Chart Average Irrelevant Links Visited 302
Figure 68 Overall Mean Completion Time All Websites 302
Figure 69 Box Plot Aggregation of Overall Time All Websites 303
Figure 70 Line Chart Killarney and Dingle Average Overall Time 303
Figure 71 Boxplot Overview Overall Time Killarney and Dingle 304
Figure 72 Average Time per Category Overall 305
Figure 73 Average Time per Category Killarney Dingle 305
Figure 74 Average Time per Category Killarney 306
Figure 75 Average Time per Category Dingle 306
Figure 76 Overall Word Count 307
Figure 77 Self-Reported Frustration Second Experiment 307
Figure 78 Self-Reported Motivation Second Experiment 308
Figure 79 Self-Reported Task Relevance Second Experiment 308
Figure 80 Influence of Browsing Skill Second Experiment 309
Figure 81 Self-Reported Importance of the Recommendations Second Experiment 309
Figure 82 Self-Reported Usefulness of System Guidance Second Experiment 310
Figure 83 Integration Overview of CSP in Open Domains Part 1 313
Figure 84 Integration Overview of CSP in Open Domains Part 2 314

List of Tables

Table 1 High-Level Design Requirements....64
Table 2 High-level requirements related to design components....73
Table 3 Closed Web Design Requirements....83
Table 4 Open Web Design Requirements....93
Table 5 Design Requirements Addressed by a Closed Web Use Environment Use case Prototype Implementation....117
Table 6 Design Requirements Addressed by CSP Prototype Implementation for Open Domains....137
Table 7 Design Requirements Addressed by the Evaluation....195
Table 8 90% Confidence Interval Total Links Visited....299
Table 9 95% Confidence Interval Total Links Visited....300
Table 10 90% Confidence Interval Average Time Dingle and Killarney....304
Table 11 95% Confidence Interval Average Time Dingle and Killarney....304

1. Introduction

"One of the effects of living with electric information is that we live habitually in a state of information overload. There's always more than you can cope with."
(Marshall McLuhan)

1.1. Motivation

The Web[1] has become an essential commodity within modern societies allowing the access, creation, management and distribution of vast amounts of information. Especially in the last decade the amount of information and the rate at which information is created has increased vastly.[2] This information explosion results in an increased difficulty in organising digital information to meet the user's information needs. To allow users to find and use information most effectively, it is essential to develop novel and innovative approaches. Several publications across different research areas, such as library and information science (Case, 2012; Korth and Silberschatz, 1997), cognitive science (Kirchhoff, 2013), social science (Chen et al., 2009) and medical science (Epstein, 2009) have voiced this need, and the failure to address it properly can result in increased costs to the economy and the wider society (Edmunds and Morris, 2000a).

In relation to the economic costs, Vivisimo[3], a leader in information optimization, estimates that the exponentially growing pool of generated content is estimated to be costing U.S. businesses $1.5 trillion in lost productivity.[4] This is a result of people spending additional time to find and organise information that is needed for a task (Feild et al., 2010).

In recent years, computer science has addressed this challenge by introducing systems that assist the user's information needs. Examples of such systems are keyword-based search

[1] The term 'Web' is used throughout the thesis as abbreviation of 'World Wide Web' or WWW.

[2] The number of blogs in 2006 alone was estimated to be 35 Million doubling every 6 months ("Sifry's Alerts: State of the Blogosphere, April 2006 Part 1: On Blogosphere Growth," n.d.).

[3] http://vivisimo.com/

[4] NewsRoomAmerica.com - Explosion of Data at Work Contributes to $1.5 Trillion in Lost Productivity. Found in http://www.newsroomamerica.com/story/49192.html (Accessed 22/04/2013)

engines (Brin and Page, 1998; Broder, 2002), browser based support mechanisms, such as bookmarking and browsing history tools (Heymann et al., 2008; Noll and Meinel, 2007), recommender systems to identify and recommend content related to the user's current interest (Adomavicius and Tuzhilin, 2005) and Web Personalisation techniques that adapt different aspects of the web experience to the needs and preferences of the individual user (Anand and Mobasher, 2005; Gao et al., 2009). Web Personalisation techniques, in particular, have become very prominent in recent years. An example of such a technique is Personalised Information Retrieval (PIR) (Speretta and Gauch, 2005) - deployed by Google[5], Yahoo[6] and Bing[7], where the search result list is re-ranked based on the user's individual search history (Speretta and Gauch, 2005). Facebook, as a second example, uses the social graph of their users to *socially connect* users with content and products that friends and peers 'like'.[8]

Even though current approaches prove to be successful in providing fast and accurate results for simple information needs, such as finding facts related to a product via a search engine or recommendation system, most applications cannot assist the user in more complex conditions. For example, if a user's information needs span different subject domains the user may need to gather information from several independently hosted websites.[9] This cross-site browsing process can introduce friction based on different website layouts the user has to adapt to in order to identify the relevancy of the navigation choices presented. Phenomena introduced in this context have become known as **lost in hyperspace** and **information overload.** Both can negatively impact task success and the

[5] http://www.google.com/

[6] http://www.yahoo.com/

[7] http://www.bing.com/

[8] Social media applications have a major impact on how individual users find and access information with several applications using the social interactions the users are providing to filter information based applications (Dieberger et al., 2000).

[9] Examples of complex information needs can be found in diverse areas, such as consumer purchasing and holiday planning, but also be found within research and commercial projects (Kellar et al., 2006a). Planning a holiday for example can take several weeks and require the access and information to several individually hosted websites such as airlines, car hire, hotels and local amenities. Purchase interests may require continuously seeking of background information and pricing and in research projects students seek and explore information about a specific topic across various different sources (Terai et al., 2008).

user's decision-making performance (Ahn et al., 2008; Berghel, 1997; Edmunds and Morris, 2000a; Hiltz and Turoff, 1985; Speier et al., 1999).

To appropriately support the user a unified and seamless browsing experience within and across different websites (e.g. wikis, blogs, forums, general content repositories, enterprise and retail websites) is required.

Current limitations in providing such support are that most websites don't offer any cross-site communication, leading to a browsing experience that is mostly confined within a website; thus if a user is gathering information about a specific topic on a website and then decides to browse to a different website, the latter will not be informed about the user's previous information needs (Abel et al., 2013; Brusilovsky and Maybury, 2002). Furthermore any personalisation techniques used by one websites will not be able to assist the user in relation a wider information needs across a second (independent) website. The effect of this silo-ing of information and personalisation techniques can lead to the repetition of user actions, such as keyword based search queries, and navigational patterns throughout the different websites, such repetitive interaction is both frustrating and time consuming for the user (Feild et al., 2010).

Even if websites were able to cross-communicate, most websites would be unable to provide effective assistance based on the individual user's vast amount of needs and preferences. To ensure effective assistance the website a user browses across will have to integrate a consistent cross-site support mechanism. This engineering challenge is currently time and cost intensive.

Furthermore, privacy implications should be addressed; the user needs to fully understand how personal information is used, collected and protected to avoid misusage. An example of this has become known as the *filter bubble*. It is based on the observation that a filtering algorithm can influence the user's opinion by filtering out information that seems irrelevant. Furthermore, even if the user wants to find out what has been filtered, only limited possibilities are offered (Pariser, 2011). Therefore, failing to provide transparency can lead to a lack of trust in personalisation approaches (O'Callaghan, 2011).

Failing to provide transparency relates to a further notable problem of Web Personalisation techniques. It lies in ensuring the user is able to freely browse without any limitations or control (O'Callaghan, 2011). This freedom however should be balanced by providing the user with non-intrusive guidance that provides assistance, but doesn't limit the user's browsing experience. In relation to this problem users are currently confronted with two interconnected web domains. The first, **closed domain** is introduced as a browsing space that is limited to a specific number of websites and a specific user group. The content is usually subject-specific and structured based on strict editorial guidelines (e.g. Enterprise level website federations). The second, **open domain**, introduces a browsing space not limited to a specific number of websites or a specific user group. The content is heterogeneous and mostly not limited to editorial guidelines or structures (e.g. blogs).

A further gap of Web Personalisation techniques is the increased engineering effort to apply Web Personalisation to existing websites. Most approaches are deeply integrated into website in the design-phase. For Web Personalisation to offer a more seamless experience across different Websites it is necessary to investigate non-invasive techniques allowing Web Personalisation to be introduced to Websites more flexibly and without any substantial engineering effort.

A key problem therefore remains in providing Web Personalisation techniques that assist users across the web and not only on isolated Websites. To address this problem Web Personalisation techniques have to balance both the needs of the website user and website publishers. For website users the introduced Web Personalisation technique should provide assistance across different websites, but in doing so honour the user's privacy needs and browsing freedom. For the website publisher the introduced Web Personalisation techniques should include simple and cost effective integration of Web Personalisation to existing websites and honour the website publisher control over the website.

1.2. Research Question

How can users be assisted in addressing information needs that span independently hosted websites; and consequently, how appropriate might that assistance be?

By appropriate it is meant *"to what extent can Web Personalisation techniques be utilised to satisfy the needs and concerns of the web user*[10] *and, if necessary, the needs and concerns of website publishers*[11]*"*.

1.3. Research Objective, Goals and Challenges

To address the above research question, the objective of this thesis is to introduce and evaluate a novel Cross-Site Personalisation (CSP) approach to assist users in addressing information needs that span independently hosted websites (Abel et al., 2013; Koidl et al., 2009; Yudelson, 2010).

The main goals of this thesis are -

- To **identify** the key requirements, techniques and impact of current Web Personalisation and Adaptive Hypermedia techniques in assisting users in addressing information needs that span independently hosted Websites.
- To **design and develop** a Cross-Site Personalisation approach to support users in addressing cross-site information needs.
- To **evaluate** the appropriateness of the developed approach based on a series of case-study prototype implementations and experiments.

To address these research goals the design and research challenges are -

- To investigate the applicability of Cross-Site Personalisation in **open and closed domains**.
- To investigate the system requirements to consider the user's **privacy concerns**.
- To investigate the **engineering effort** on the underlying website to ensure minimal integration effort of Cross-Site Personalisation into websites.

[10] In this thesis a user is referred to as a website consumer with the purpose or goal to address information needs.

[11] By website publisher this thesis refers to a person or organisation that controls websites and therefore provides means to enable assistance.

1.4. Methodology

This research focuses on introducing and evaluating a Web Personalisation approach that assists users in addressing information needs that span independently hosted websites.

This is achieved through a sequence of exploratory and formative research studies. For exploration the type of research methods applied is descriptive and executed through user-focused surveys, fact-finding enquiries and task-based evaluations. For confirmation comparative and correlative methods are used.

The conducted research can be categorised as **applied research** based on the investigation and development of a novel Web Personalisation approach that is applied to websites with the goal to assist users in addressing information needs that span independently hosted websites.

Qualitative and quantitative research techniques were applied to address the research question of this thesis. **Qualitative research** techniques were applied to understand the underlying motivation of how and why users relate to the topic of information overload, Web Personalisation and online privacy. Based on focused and semi-structured interviews, attitudes, opinions and behaviour in relation to the research question were identified. This led to the design and development of a novel Web Personalisation approach, assisting users in addressing information needs that span individually hosted websites. **Quantitative research** techniques were applied to study the effect the introduced Web Personalisation approach has on the user addressing cross-site information needs. The underlying approach for these quantitative studies are experimental by controlling the conditions through task-based assignments in pre-set browsing environments based on real-world use cases and real-world content.

1.5. Contribution

The contribution of this thesis is the introduction of a Web Personalisation approach focused on assisting the user in addressing information needs that span independently hosted websites. This is achieved by introducing a third-party personalisation service together with website extensions ensuring non-intrusive and privacy sensitive assistance

across the web. Moreover this approach is a novel engineering contribution to Web Personalisation by ensuring the integration of Cross-Site Personalisation in existing websites is as simple and cost effective as possible (Koidl and Conlan, 2008).

A minor contribution consists of studying, highlighting and addressing the challenges of evaluating Web Personalisation techniques with limited resources and budget. This challenge is addressed through a series of case-based evaluations with different user groups and content types. The evaluations indicate a positive impact of Cross-Site Personalisation based on the user-performance results and the self-reported feedback provided by the participants. Moreover the evaluation results demonstrate that the proposed approach in assisting users in addressing information needs across independently hosted websites is appropriate in both open and closed domains.

This investigation has resulted in the following publications:

Koidl K., Conlan C., Wei L. and Saxton A.M. (2011) Non-invasive Browser Based User Modeling Towards Semantically Enhanced Personalisation of the Open Web. FINA-3A: Semantic Web and Systems, AINA 2011, Singapore.

Koidl K., Conlan O., Wade V. (2009) Non-Invasive Adaptation Service for Web-based Content Management Systems (AH2009), International Workshop on Dynamic and Adaptive Hypertext: Generic Frameworks, Approaches and Techniques, Torino, Italy.

Koidl, K., Conlan, O. (2008) Engineering Information Systems towards facilitating Scrutable and Configurable Adaptation, Fifth International Conference on Adaptive Hypermedia and Adaptive Web-based Systems (AH2008), Springer Lecture Notes in Computer Science 5149, Hannover, Germany, 28 July - 1 August 2008, pp405-409.

This research was the basis for three completed MSc projects. Furthermore the outcome of this research has been filed as full patent application (US and UK) and has resulted in three successful funding proposals to investigate the commercial impact of the approach.

1.6. Scope

The scope of the thesis is focused on assisting users in areas related to information overload, such as navigational guidance, online privacy and user model scrutiny. Closely related research areas are Adaptive Hypermedia Systems, Information Retrieval and Recommender Systems.

The overarching motivation is to assist users in addressing cross-site information needs by focusing on identifying and understanding information needs that require browsing independently hosted websites.

In relation to Adaptive Hypermedia (AH) the thesis focuses on heterogeneous content sources, decentralisation, user modelling and the open-corpus problem (Brusilovsky and Henze, 2007). Furthermore AH techniques, such as link annotation are discussed and applied. Related and studied research areas are Recommender Systems and Information Retrieval.

To allow the semi-automatic and automatic extraction of topics and terms from web content text analytics techniques are discussed and applied.[12]

To investigate the engineering challenge of assisting users across independently hosted websites, Web-based Content Management Systems (WCMS) were identified as initial proof of concept platform.

1.7.Thesis Overview

The thesis starts with the State of the Art review (chapter two) discussing different research and technological approaches related to Cross-Site Personalisation. The State of the Art review is focused on identifying and discussing challenges in the research area of Web Personalisation. This discussion results in the structuring of the remainder of the State of the Art in three interconnected web dimensions of Cross-Site Personalisation: (a) **The Web User Dimension**, focusing on identifying and understanding the user's information needs,

[12] Creating a unified abstraction layer across the different content types hosted on the different websites the user browses across can be seen as a main challenge for this research. Challenges include the need for a unified vocabulary, which also relates to the research fields of ontology mapping and taxonomy creation.

how information needs can be modelled and what impact privacy has on modelling information needs; (b) **The Web Content Dimension**, focusing on Web Personalisation across different web content types and domains; (c) **The Web Application Dimension**, discussing Web Personalisation systems and architectures that facilitate Web Personalisation. The focus of the application dimension is on non-intrusive Web Personalisation techniques, interoperability (portability of models and/or other elements of the approach) and the non-invasive integration of Web Personalisation in existing websites. The latter investigates web system architectures, such as WCMS to identify and reflect web-engineering challenges.

The State of the Art discussion is followed by an investigation of Cross-Site Personalisation use cases in open and closed domains. The investigation is based on an initial user survey and subsequent web-based analysis of subject domains and websites.

The investigation of cross-site use cases based on user feedback resulted in the design (chapter three) and implementation (chapter four) of **two use case based prototypes**. The first prototype focuses on applying Cross-Site Personalisation within **closed domains**. Based on the outcome of the first prototype, a second prototype was developed. This prototype illustrates the application of Cross-Site Personalisation within **open domains**. Chapter five of this thesis evaluates both prototypes in two user-focused evaluations. The final chapter (chapter six) discusses conclusions and further work related to Cross-Site Personalisation.

2. State of The Art Review

"The web is continuously transforming itself from providing a 'one size fits all' experience across all users to a space within which people can carve out their own niches"

(Berners-Lee et al., 2006)

2.1. Introduction

This chapter discusses the State of the Art in Web Personalisation techniques and methods with the objective to identify challenges for assisting users in addressing cross-site information needs.

Web personalisation can be defined as any action that tailors the web experience to a particular user, or set of users, by providing the information users want or need without expecting them to ask for it explicitly (Keenoy and Levene, 2005; Mobasher et al., 2000; Mulvenna et al., 2000). Methods and techniques enabling Web Personalisation have become important for both research and commercial applications (Liewehr and Laplante, 2011; Mobasher et al., 2000). The main motivation of Web Personalisation is to assist the user in managing the access to vast amounts of rapidly growing information spaces (Belkin and Croft, 1992; Nanas et al., 2009). This development, often referred to as *information explosion* (Rudd and Rudd, 1986; Sweeney et al., 2001), has led to an increase of psychological phenomena's known as *information overload*[13] (Berghel, 1997; Edmunds and Morris, 2000b) and *lost in hyperspace* (Edwards and Hardman, 1999). The former referring to the user being overwhelmed by the amount of information choices and the latter referring to orientation difficulties of the user within and across different websites.

This thesis argues that in order to assist the user across the web it is important to develop Web Personalisation techniques and methods that assist the user in addressing information needs seamlessly across independent websites.

[13] Information Overload is strongly connected to the concept of *cognitive overload*, a state in which the working memory can overload resulting in a decreased capacity for the user to solve a task or address a need (Sweller, 1988).

Adaptive Hypermedia (AH) can be noted as one of the main research fields addressing the challenge of assisting the user addressing in information needs. Especially in domain specific application, such as e-Learning, the use of AH techniques and methods have been proven to be successful by increasing the learning performance of the student (Bailey et al., 2002; Brusilovsky, 2001; Conlan and Wade, 2004). However, this domain focus has also lead to a shortcoming known as the open-corpus problem and described as: '*The problem to provide adaptation within a set of documents that is not known at design time and, moreover, can constantly change and expand*' (Brusilovsky and Henze, 2007).

This thesis seeks to extend the open-corpus problem of AH by introducing the **open-web problem of personalisation***:* The *open-web problem of personalisation* extends the open-corpus problem of Adaptive Hypermedia by adding '[…] *and which may reside in individually hosted websites on the open web*.'

To address the open-web problem of personalisation this thesis introduces the concept of C*ross-Site Personalisation (CSP)* as a process in which a web user is individually assisted in addressing information needs that span independently hosted websites.

In the context of this thesis information needs that spans independently hosted websites are referred to as *cross-site information needs*: A perceived lack of information that the user requires to complete a task or intent or to pursue an interest that requires browsing several independently hosted websites.

To study and investigate the challenges and requirements of CSP the following interconnected dimensions are defined: (1) The **Web User Dimension,** reflecting the needs and concerns of the web user; (2) The **Web Content Dimension**, reflecting the information a user requires to address information needs; (3) The **Web Application Dimension,** reflecting the needs and concerns of the website publisher. Figure 1 illustrates the interconnection between all three dimensions enabling a fourth dimension, the **Web Personalisation Dimension** with the goal to balance the needs and preferences of the users (user dimension) and the website publishers (web application dimension) by establishing a state of equilibrium (introduced as *Equilibrium of Web Personalisation*).

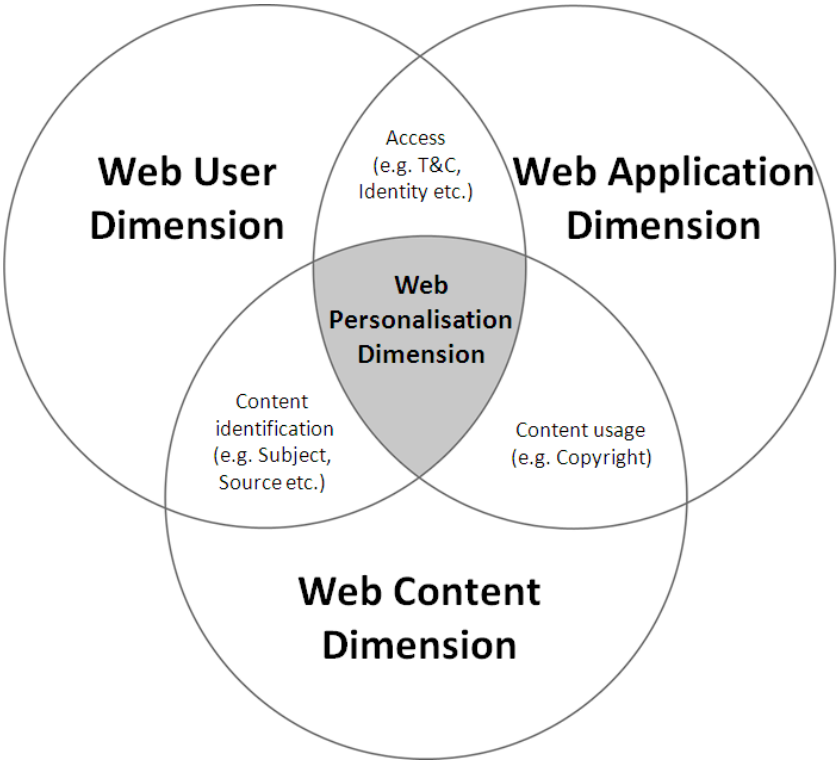


Figure 1 Web Personalisation Dimensions

This State of the Art discussion focuses on identifying and discussing the challenges and requirements of Web Personalisation at the intersection of the above described dimensions.

The remainder of the State of the Art chapter is structured as follows: (1) A brief discussion about information overload and the limitations of current search paradigm with the goal to motivate the higher level problem space addressed by CSP, (2) a brief discussion of the identified gap in the State of the Art, and (3) a subsequent State of the Art discussion of each the three dimensions. The outcome of the State of the Art discussion is the identification of main challenges and requirements to be addressed by the design, implementation and evaluation of a CSP approach.

2.2. Information Overload and the Limitations of the Current Search Paradigm

It has become increasingly difficult for information systems to effectively assist the user in addressing information needs. The main reason is an abundance of data due to advancements in web technologies that allow users to easily create, share and publish information. One side effect of this information explosion has become known as *information overload*[14], which can result in disorientation and the risk of misinterpretation of information (Chen and Sycara, 1998).

To address the effects of information overload various information access technologies have been introduced, including *Information Retrieval* (Salton and McGill, 1986) following the *needle in the haystack* notion, *Text Mining* (Tan, 1999) to extract meaningful information and behaviour patterns from large information spaces, *Recommendation Systems* (Adomavicius and Tuzhilin, 2005) using mostly user data correlation techniques to recommend relevant content, and finally *Adaptive Hypermedia* (Brusilovsky, 2001; Keenoy and Levene, 2005) providing adaptive assistance based on the user's needs and preferences.

It can be argued that currently the most applied approach in addressing information overload on the web is Information Retrieval. This approach is highly effective in addressing the user's partial information needs, which is expressed in a sequence of keywords provided by the user.

---

[14] Based on the above description this thesis follows the notion that information overload is a state in which a user is overwhelmed with the amount of information presented resulting in a lack of judgement in what information is relevant. Information overload can be defined as a cognitive process in which the user is overwhelmed with the amount of accessible information. This phenomena has been known since the early 50s when George Simmel, a social scientist, mentioned a sensation in the urban world that caused city dwellers to become tired and develop a incapacity to react to new situations with appropriate energy (MacRae and Wolff, 1951). More recent definitions related to information overload can be found in (Stanley and Clipsham, 1997), describing the phenomena in terms of analysis paralysis and information fatigue. Feather, as another example, describes information overload as a point where there is too much information, which results in a situation in which it cannot be used effectively (Lievrouw, 1996). More recently information overload has been brought into connection with research related to *cognitive load* which relates to the working and short-term memory. Specifically research in the identification and measurement of *mental workload* has introduced several theories to measure information overload and information underload. For a comprehensive literature review note (Cain, 2007).

The current search paradigm has three limitations: (1) The user's expression of need is based on the interaction within the website of a search engine. It is difficult for a search engine to assist a user outside of the confines of the search engines website; (2) The search query of the user is only a partial representation of the user's overall information needs; (3) Search engine technology can lead to a filtering effect that narrows the choice and informational self-determination of the user (O'Callaghan, 2011).

In addition to search related approaches to information overload, browsing related techniques have been introduced. The main focus of browsing assistance mechanisms is to guide the user to the right content through manipulating the websites link structure (Brusilovsky, 2001; Teevan et al., 2004).[15] Typical approaches are known as *link highlighting* and *content recommendation,* which range from a simple colour overlay to more intrusive techniques, such as reordering and hiding mechanisms (Brusilovsky, 2008; Pazzani and Billsus, 2007). The main shortcoming of these techniques is that they are deeply integrated in the application making it difficult to apply across different applications. This shortcoming becomes relevant in cases the user addresses overarching information needs across several different applications and/or websites and may result in increased disorientation and user frustration due to isolated personalisation's (Feild et al., 2010). In addition to isolated application based approaches, browser or device-based techniques have been introduced (Perugini and Ramakrishnan, 2003).

This thesis argues that the current search paradigm provided by search engines, as well as the current isolation of personalisation techniques, such as provided through Adaptive Hypermedia Systems and Recommender Systems, have not sufficiently addressed the challenge of information overload in information needs that span independently hosted websites. To assist the user in cross-site information needs several different challenges need to be identified and addressed.

[15] Navigation in hyperspace has been referred to as 'orienteering' in comparison to search which has been mentioned as 'teleporting'. The latter is based on the notion that the user directly lands on the page of interest and does not need to navigate towards it (Teevan et al., 2004).

2.3. Gap in the State of the Art

Web personalisation, as stated in the introduction of this chapter, can be defined as a process that allows tailoring of the user's web experience to assist the user in addressing information needs (Keenoy and Levene, 2005). The main objective of Web Personalisation therefore is to influence the interaction between the website and the user through means of information access and presentation with the goal to assist the user in experiencing the most relevant content. Based on this goal, Web Personalisation has become popular in mass user web applications, such as social media (Stewart et al., 2009), retail (Schafer et al., 1999) and search platforms (Keenoy and Levene, 2005; Pazzani and Billsus, 2007).[16]

Even though Web Personalisation has reached a high-level of maturity it still has several technical and non-technical challenges (Kim, 2002; Liewehr and Laplante, 2011; Shahabi and Chen, 2003). The main challenges are related to user identification (Carmagnola and Cena, 2009; Pazzani and Billsus, 1997; Yinghui, 2010), understanding user behaviour (Berendt and Spiliopoulou, 2000; Yinghui, 2010; Žumer, 2012), user modelling (Carmagnola et al., 2011; Dolog and Nejdl, 2003; Stadnyk and Kass, 1992), entity extraction (Johnson and Kumar Gupta, 2012; Mobasher et al., 2000), content modelling (Oard and Kim, 2001), as well as web engineering challenges (Ginige and Murugesan, 2001; Koch et al., 2008), such as integration of Web Personalisation techniques and finally privacy concerns (Kim, 2002; Lee and Cranage, 2011).

This thesis identifies and addresses **a gap in the State of the Art** related to Web Personalisation. Currently Web Personalisation techniques provide an isolated experience only effective on the current website the user is browsing, but not across different websites. Reasons for this can be found in the continuing isolation of web experience referred to as

[16] Within these applications mostly two types Web Personalisation techniques are applied: (1) Information filtering techniques, such as collaborative and content-based recommendations (Schafer et al., 2007) (Pazzani and Billsus, 2007), and (2) adaptive hypermedia techniques, such as link or content highlighting (Brusilovsky and Maybury, 2002). Further techniques are related to web usage mining (Mobasher et al., 2000), user tracking (Oberle et al., 2003) and online privacy (Adkinson et al., 2002) (Lee and Cranage, 2011). Furthermore, the recent mass adaption of mobile devices has added an addition facet to Web Personalisation by providing experiences based on location awareness and device functionality (Korpipää et al., 2004). Especially related to the limited display capability of mobile devices, Web Personalisation techniques can prove essential for successfully assisting user within mobile application.

walled gardens (Berners-Lee, 2010; Koidl et al., 2009). This isolated Web Personalisation experience can result in a fractured browsing experience possibly disrupting the overall browsing process especially if the user is browsing the web with overarching information needs. As motivated in the introduction of this section, this shortcoming relates to the open-corpus problem identified by various research groups (Bailey et al., 2002; Brusilovsky and Henze, 2007; Brusilovsky and Maybury, 2002; Henze and Nejdl, 2001).

This thesis seeks to address this gap in the State of the Art by designing, implementing and evaluating a CSP approach that assists users in addressing information needs across individually hosted websites.

In the following three sections are structured based on the three interconnected dimensions introduced in Figure 1 (2.1).

2.4. Web User Dimension

"What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently among the overabundance of information sources that might consume it."
(Herb Simon)

2.4.1. Introduction

Identifying, understanding and subsequently assisting the web user in addressing information needs is a wide and complex undertaking and far outreaches the scope of this thesis. However, in the following, a focused attempt is made to highlight and discuss the current State of the Art of understanding the user's information needs.

For this, the discussion is split into three separate sections. The first section discusses approaches in understanding and assisting the web users in addressing information needs. The second section focuses on modelling information needs. Finally, the third section discusses approaches reflecting user identification, online privacy and trust.

2.4.2. Understanding and Assisting the Web Users in Addressing Information Needs

To assist the user across independently hosted websites, it is important to identify and understand the information needs of the web user. This section will start by briefly discussing *what* information need are and *how* these originate. This initial discussion is followed by a discussion about how the user's information needs are identified followed by a discussion on approaches that assist the user in addressing information needs.[17]

What are information needs and how do they originate?

[17] It should be noted that the understanding of the web user is a complex and comprehensively researched topic that extends across several different research areas, such as human computer interaction, web science and cognitive science. Therefore the following investigation is not complete. To date no unified theory or model has been introduced (Dix, 2004).

Research in investigating the understanding of the user's information needs can be traced back to the early 60s. Then library catalogues moved from purely paper card based item-search to computer supported topic-search (Licklider et al., 1968). This fundamental shift in how a user expresses information needs, not by searching for an item, but searching for topic, opened the door to a wider discussion in both the source of this need and how this need can be assisted most effective.

Mackay (1960) defines information needs as: *"[…] certain incompleteness in his [the inquirer's] picture of the world – an inadequacy in what we may call his 'state of readiness' to interact purposefully with the world around him in terms of a particular interest"*. A focused investigation of the user's information needs was conducted by Taylor (Licklider et al., 1968) resulting in the following four types of information needs: (1) *visceral needs* as actual, but *unexpressed need* for information; (2) *conscious needs* as a within-brain description of a need; (3) *formalised needs* as a formal statement of information needs and finally (4) *compromised needs* as a question presented to the informational system. Following the investigation of Taylor, Belkin (Belkin et al., 1982) studied the origin of information needs further. He defines information needs as a perceived lack or problem that the user wants to address or solve. Belkin's main contribution is to view information needs as a problem identification process that leads to an *Anomalous Knowledge State (AKS)*. The ASK is described as a state that arises if the user's state of knowledge about a topic is inadequate and the user thinks to be missing the ability to solve a problematic situation. The importance of Taylors and Belkin's work in relation to this research lies in introducing the understanding that the user mostly addresses information needs through a *sequence* or *process* and not one single interaction, query or request. In a later stage the notion of a process has been argued by Dervin (2012) in relation to sense making. Here, a user seeks to bridge the AKS by seeking for information[18] that makes sense in relation to information needs. Once relevant information has been found the user moves forward to the next informational item until the need is fully addressed (Hellerstein et al., 1996).

How to identify the user's cross-site information needs

[18] It should be noted that Dervin does not see information as an individual object or entity, but as embedded information in an information object that should be conceptualized by the user in a particular situation.

The activities the user conducts to address information needs is mostly based on a process that includes a mixture of searching and browsing (Olston and Chi, 2003). Even though browsing is an essential part of this process most research groups have focused on identifying the user's information needs within web-based search engines. The main reason for this lies in the user expressing their need in a sequence of explicitly stated keywords. Identifying the user's needs in browsing activities is far more complex. The two main reasons are: Firstly, the user's expression of need is implicit, such as based on the time users spend on specific pages, what patterns the users navigate and any other track able actions, such as keyboard presses and mouse clicks. Secondly, and more significantly, websites are mostly isolated from each other, not facilitating a centralised storage of the user's interactions; thus making the identification of user needs a complex undertaking.[19]

User need identification is usually divided into the two areas of explicit feedback and implicit feedback. Examples of explicit feedback are user stated preferences and ratings, such as based on a 5-star rating, thumbs up or down and social media buttons (e.g. 'Like' and '+1'). It offers simple means to gain a precise understanding of the user's needs and preferences. However, three shortcomings can be identified: (1) The user is required to interrupt the current task to provide a rating. In some cases this even requires signing into a service or website. (2) Users don't rate everything they perceive interesting. Finally, (3) an interest stated through explicit rating may only be relevant and accumulate the moment the users provides the rating and not at any later stage.

Implicit feedback provides a good alternative to these shortcomings by relying on less intrusive measures, such as reading time and clicks. In relation to accuracy research conducted by Claypool et al. (2001) indicated no significant difference in the accuracy of both feedback mechanisms.[20] A further challenge of implicit feedback is the identification if

---

[19] A research field addressing the analysis and understanding of the user's browsing patterns has become known as web usage mining. Here, log analysis provides various interesting outcomes, such as identifying users based on browsing patterns (Mobasher et al., 2000) or using the stored log information for Web Personalisation. However, based on the disparity of websites especially if not related and/or independently hosted, this technique is only useful within the confines of one website and not across websites.

[20]

the user is viewing the page even though no action (e.g. clicking or scrolling) is performed (Morita and Shinoda, 1994; Resnick et al., 1994). Identifying viewing and idle time may be important to identify how relevant the content of the page is. There is no clear research on how long idle time should be before concluding that the user is not actively reading the content. An approximation can be found in analytical tools were 30 minutes (1800 Seconds) idle time on a web page is usually considered as a threshold (Google Analytics, 2012; Ledford et al., 2010).

After discussing *how to identify* information needs the following discussion focuses on *how to assist* the user.

How to assist the user in cross-site information needs

Early research in assisting the user co-insides with Taylor's early work on identifying and classifying information needs (Licklider et al., 1968). The interesting notion introduced by Taylor was the introduction of trained human intermediary that ensured the user's needs are correctly identified, understood and assisted. For this, Taylor defined five steps of interaction between the information seeker and the intermediary defined as *filters* (Licklider et al., 1968)

1. Identify the subject of information needs.
2. Identify the objective and motivation behind information needs.
3. Identify various personal characteristics of the subject who has information needs.
4. Establish the relationship between the description of information needs, the file organisation and the language.
5. Determine expected or acceptable answers.

It seems plausible that the length of the page has a strong correlation on the reading time. However, research conducted by Morita and Shinoda (1994) indicate that there is no significant correlation between the time a user spends on an interesting page and the length of the page with an average reading time on interesting articles of 20 seconds. Examinations conducted by White et al. (2002) in relation to reading time and page length were also inconclusive. Identifying the main indicators expressing interest based on implicit feedback therefore should have a focus on time on page. Considering time on page however should take user actions into account, such as scrolling, clicking and key presses to ensure only active reading time is taken into account. Hauger et al. (2011) identified a strong correlation between the gaze of a user and the mouse pointer position which indicates that for the identification of relevant text passages mouse interaction may be useful.

The importance of Taylors *filters* in relation to this research is the process-based approach to understanding how the user wants to be assisted by an intermediary.[21]

Following Taylor's early pioneering work many research fields have concentrated on the development of an intelligent intermediary to assist users in addressing information needs. However, even with this considerate effort, it still remains difficult to successfully assist a user in addressing information needs, especially if they span separate information sources.

One of the main reasons for this difficulty is to find the right balance between *helping* and *getting in the way*. Especially in Web Personalisation ensuring this balance can be the difference in the user feeling *lost* or *controlled*, *overwhelmed* with choices or *guided* towards more of the same.

One of the most prominent research projects investigating this balance was the Lumière Project (Horvitz et al., 1998). The project resulted in the offline office application assistant known as 'ClipIt' or 'Clippy'. The mechanism uses a sophisticated probabilistic Bayesian user modelling approach to reason about the user's time varying goals and needs throughout the application. Even though comprehensive research has been conducted it can be concluded that the project failed due to the user perceiving the interaction with the assistant as negative and highly intrusive (J. Xiao et al., 2003). The important contribution coming from the Lumière Project was in identifying the challenges in understanding and assisting the user. The main challenges identified can be summarised as: (i) Discovering the type of

[21] Related research by Taylor (1967) and Parker and Paisley(Parker and Paisley, 1966) also indicated that a user seeking information tend to consume the incorporate relevant information into a cognitive state or model. Related to modern web systems these observations motivate the difficulty in applying assisting technologies such as personalisation due to not allowing the user to build this knowledge map them self, but more allowing a user to build this map. Following the notion of a conceptual map later research such as by Belkin (1980) argues that a user is able to detect an anomaly in the state of knowledge, which hinders the user to accomplish a specific goal or purpose. In relation to the overall task identification of the user this notion becomes essential in allowing the user understand the connection of individual browsing activities that build the user cognitive map i.e. if a user browses across different websites it is essential for the user to connect relevant concepts from those websites in an overall cognitive map. If this map gets disrupted or if the user detects an anomaly in the state of knowledge such as introduced by Belkin, the user will seek to overcome this anomaly or suffer disorientation and information overload.

assistance a user wants and (ii) learning when (and if) to assist the user (Schiaffino and Amandi, 2004).

Assisting the user the right way has been extensively researched in the context of Adaptive Hypermedia Systems (AHS) (Brusilovsky, 2001, 1996). The two main support mechanisms introduced in AHS are *adaptive presentation* and *adaptive navigation support*. The following Figure 2 illustrates the different support mechanisms applied in AHS.

Figure 2 Taxonomy of Adaptive Hypermedia assistance mechanisms (Brusilovsky, 2001, 1996)

Assisting the user by applying adaptive navigation support introduces the challenge of identifying when the user requires assistance and how intrusive this assistance should be. In a navigational setting, such as the web, this challenge can be addressed by trying to identify if the user is lost. However, detecting disorientation in user navigation is highly complex (Juvina and Herder, 2005). The main reason for this can be found in the individual perceptions and navigations style of the user. Purely by using idle time it is not possible to indicate that the user is disoriented or lost. Furthermore the user's individual abilities and

domain expertise impacts the user's sense of disorientation. It is also not obvious from a systems standpoint if the user is just *snooping around*, *spending time,* hoping for *serendipitous finding* or if the user is actually trying to fulfil information needs.[22]

This challenge has been investigated by Ahuja and Webster (2001) through the design of a user questionnaire that helps in use the pre-stated preferences of a user to identify disorientation. A more dynamic approach has been taken by using the amount of revisits to specific webpage as an indicator of disorientation (Herder and van Dijk, 2005). Some studies also indicate weak navigation styles (Herder and Juvina, 2004) and backtracking in browser usability (Catledge and Pitkow, 1995) as source of disorientation. A further interesting approach is introduced by Smith (1996) by introducing the concept of *lostness*. Lostness is calculated using three values: (N) The number of different web pages visited, (S) the total number of pages visited including revisits and (R) the minimum number of pages required to fulfil the task. Lostness is calculated by using the following formula:

$$L = \sqrt{\left(\frac{N}{S} - 1\right)^2 + \left(\frac{R}{N} - 1\right)^2}$$

A score above 0.5 may indicate that the user is lost. Even though the concept of lostness provides simple means to calculate the level of disorientation (and compare the score with other users) its main shortcoming lies in the need to identify the exact number of pages necessary to fulfil a task. Therefore lostness may be useful in experimental settings, but its applicability in open web, and more explorative use cases, may be difficult.

In the following the State of the Art in modelling the identified user needs and preferences is discussed. The focus herein lies in discussing challenges in modelling the user's information needs.

[22] It can be argued that both techniques can be applied in different phases. Adaptive navigation support provides assistance to the user in accessing relevant content through techniques, such as link annotation, and adaptive presentation is useful when the user has accessed the relevant page. Here techniques, such as the highlighting text segments that are most relevant can be applied. This however implies that the assistance mechanism has a deep understanding of the different segments within a webpage.

2.4.3. Modelling Information Needs

User modelling is a wide and complex research area. The following discussion is focused on identifying and addressing challenges in current State of the Art in assisting users across separate websites. The following topics are discussed: User model types, implicitly modelling the user across separate websites, cold start problem, modelling temporal aspects, model gaps and addressing separate needs

User Model Types

User models are often implemented as an overlay model based on or *laid over* the concepts of the application (Brusilovsky et al., 2005a). The main benefit is increased accuracy and performance. The shortcoming, however, is a limitation of the models interoperability within other applications or subject domains (Carmagnola et al., 2011).

User models that allow a higher-level abstraction are often referred to as user profiles. User profiles can be defined as a subclass of user modeling with user profiles being less sophisticated and more suited for applications that require a more general abstraction of information needs (Koch and Wirsing, 2001). User profiles can be based on implicit or explicit data. Due to the browsing notion within CSP the following discussion focuses on implicit user data. The three main user profile types discussed are keyword-based profiles, semantic network profiles and concept profiles (Gauch et al., 2007).

Keyword based profiles

Keyword based profiles are based on keywords extracted from web pages, such as tags, metadata and keywords explicitly provided by the user. This type of profile typically consists of keywords from websites associated with a form of weighting typically ranging from 0 to 1. Keyword based profiles are usually based on a single term vector and use a TF-ID approach to extract terms from content (Salton and McGill, 1986). Examples of approaches using TF-ID based keyword extraction are Amalthaea (Moukas and Maes, 1998) and Webmate (Chen and Sycara, 1998).

The main drawback of keyword-based profiles is the lack of understanding in the meaning of the stored terms. This problem is known as the *polysemy problem*. For example, the term

bank can relate to a *riverbank* or a *financial institution*. Approaches extracting keywords based on information retrieval mechanisms, such as TF-IDF, are strongly affected by this shortcoming (Meadow, 1992). PSUN (Sorensen and McElligott, 1995), as one example, addresses this shortcoming by using weighted n-grams allowing the user to gain a better understanding of the terms meaning. Alipes (Widyantoro et al., 1999), as a second example, assigns three keyword vectors to a higher-level interest. Finally, to extract the higher level interest implicitly PEA (Montebello et al., 1998) uses the user's bookmarks by indicating a bookmark as a higher level interest and associating a term vector to the bookmark.

Semantic Network Profiles

To allow the user to gain a deeper understanding in the meaning of the extracted terms and to overcome the problem of polysemy semantic term networks can be used. Semantic networks usually consist of a term structure, which entails an order or relationship. Based on the example above, the term *Bank* relates to the term pair (n-gram) *Financial Institute*. This allows a system to understand that the term *Bank* does not relate to a riverbank. A further advantage of Semantic Network Profiles over keyword-based profiles is that the connection between the keywords and the higher-level concept can be maintained. This leads to a graph like structure in which the nodes represent keywords or concepts and the arcs connections. Following the example above the concept node *Financial Institute* may have several different keyword nodes associated with it, such as *bank*, *coffer* and *credit union*. The main shortcoming of creating Semantic Network Profiles is that they can be time consuming to create and keep up to date.

Concept Profiles

Concept profiles are similar to term based profile with the difference that higher level topics are displayed, which allows the user to gain a better overview of the interest the system has inferred. For example, a concept profile would indicate to the user that an interest was identified related to higher-level n-gram *Financial Institute,* but not display the lower level term *Bank*. This higher-level topic extraction allows a profile to become more focused without showing the user all terms, such as in a purely term based profile. A different

example of a concept-based profile is based on a thesaurus. For example, Wordnet[23] allows the extraction of term relationships and synonyms to facilitate the construction of a concept profile. Open Directory Project (ODP)[24] (White et al., 2009) as a second example, allows the extraction of a yellow page based hierarchy of terms. The main shortcoming of concept profiles is that usually rely on taxonomy or ontology[25], which are time and cost consuming to create and update.

In the following further identified challenges modelling the user's cross-site information needs are discussed. These are modelling the user across separate websites, cold start problem, modelling temporal aspects, model gaps and modelling different types of information needs.

Modelling the user across separate websites

In relation to CSP it can be argued that modelling the user's needs and preferences across independent websites creates a type of conundrum, in which it is necessary to assist the user not only in one website, but across several websites with different subject domains. To address this challenge a possible approach is the introduction of a *shared conceptualisation* of the overall browsing space of the user. A shared conceptualisation could consist of an overarching vocabulary reflecting the user's browsing space. This would allow the user to receive a constant and unifying browsing experience towards addressing cross-site information needs (Gruber, 1993).[26]

Cold Start Problem

[23] http://wordnet.princeton.edu/

[24] http://www.dmoz.org/

[25] A taxonomy can be defined as a concept or term structure (Maedche and Staab, 2001). A more advanced concept to represent relationships of terms is an ontology, which, in addition to only relying on the structure i.e. connection between the concept and terms, adds relationships to allow logical reasoning.

[26] The terminology of a shared conceptualisation is often used in ontology based research and the semantic web. In the context of this research a conceptualisation is referred to as an open and evolving representation of the user's information need across the web. Ontology (or semantic web) based approaches usually rely on a pre-set structure or standard, such as RDF (Dolog and Nejdl, 2003).

A further challenge in the context of modelling user needs is known as the *cold start (or new user) problem*. This problem relates to the necessity of bootstrapping the model of new users (Fischer, 2001). In e-Learning applications the cold start problem is usually addressed by asking the user to fill out a questionnaire in which the user provides initial information that populates the model. Recently social media has been introduced as an alternative solution to the cold start problem (Abel et al., 2010). It can be argued that both approaches are not ideal for browsing that may include spontaneous and unpredictable actions. A questionnaire is usually tied to the specific application. The latter, social media based solution, bears the challenge of figuring out what information provided in the social stream of the user is relevant.[27]

A third possible approach to the cold start problem is the use of collaborative approaches known as item-to-item collaborative filtering. Here the algorithm tries to correlate the navigational behaviour of groups of users. This approach is very successful with retail website where the application uses the purchase information of their users to correlate purchase tendencies. This allows recommendations, such as 'customers that bought x also bought y' (Linden et al., 2003). However, most approaches target a specific need that has been expressed through an explicit interaction, such as a purchase. If this interaction is missing, the cold-start problem may occur.

Modelling Temporal Aspects

As discussed in the previous section, the main challenge implicit user profile creation faces is to ensure the information of the user is described accurate and up to date (Stadnyk and Kass, 1992). A negative example of missing temporal aspects is a system using a user profile that represents a need that has passed, such as assisting a user in exploring holiday locations after the holiday is over.

[27] A balancing approach has been introduced by the social platform Facebook[27] by offering a social plugins[27] making a user's friends social activity available through an API. For example, it allows the consuming application to scan the social graph of the user and indicate what web pages of the websites have been liked by friends and peers within the social graph. This social approach however underlies the assumption that the interests the user's friends and peers have match the current intent of the user. Furthermore, it can be argued that the indication of *'like'* on a link by someone within the user social graph may only reflect a current mood and not a genuine interest or intent.

To address temporal aspects two different categorisations can be introduced: (1) *Indefinite* information needs and (2) *definite* information needs. The former relates to needs for which a start and end cannot be defined. The latter relates to interests with a definite start and end (e.g. exploring possible presents for an upcoming birthday). The main difference between both types is that there is no factual evidence of when indefinite information needs starts or ends. It can only be argued that the intention in addressing this need will vary over time. Addressing this aspect may require the introduction of a *deprecation factor,* which can be used to either increase or decrease the impact of current need within the user's profile. An example system using a deprecation factor is Letizia (Lieberman, 1995). Here the idle time is used to gauge an understanding if the user is not sure what to do next or if the user is simply not interested. To address this challenge it may be necessary to enable the user to provide more explicit feedback on temporal aspects of information needs.

Model Gaps

An additional challenge in relation to modeling user needs across websites is related to modeling possible gaps in the model. For example, should the user, e.g. due to privacy reasons, seek to not use the assistance on every website, the understanding of the user's information needs may be lacking accuracy due to the gap in the user's model. A research field discussing related topics is user model scrutiny (Kay, 2006). Even though this research field is focused on offering scrutiny related to the collected and modeled user information it opens the discussion to a wider user participation in the collection and usage of collected user data.

Modelling different types of information needs

The user might choose to switch needs within one browsing session. A cross-site user model therefore requires a good understanding of the user's browsing behaviour in order to model sudden switches in the need of the user. This understanding can either result from explicit user control (e.g. the user informs the assistance mechanism that different information needs are addressed) or implicitly that may require in depth understanding about the user's browsing behaviour and the user's browsing space (e.g. should the user continuously

browse on websites related to holiday location and switch to a website related to office appliances the assistance mechanism may carefully assume that the user is addressing a new/different need).

User modelling, especially on the web, holds varies additional challenges, which are outside of the scope of this research. In the following the focus lies on user identification, privacy and trust. In the following these topics are discussed.

2.4.4. User Identification, Online Privacy and Trust

User identification is an important element of information based computer based systems. Generally user identification is provided through a simple username/password request that is only known by an authorised user or user group (Carmagnola and Cena, 2009; Yinghui, 2010). By identifying the user the system can be adapted to the user profile, which among others includes right and access management. User identification has been introduced across platforms. Examples are among others *openID*[28] and *MS passport*[29]. Recently user identification has been introduced by social platforms (Miculan and Urban, 2011). This mechanism allows users to log into different applications with the credentials used within social platforms, such as the social networks Twitter [30]and Facebook[31].

Both mechanisms allow a single-on to access different applications, such as online mail and office applications. A more implicit mechanism to identify the user is through browsing session and cookie detection. A cookie is a small file that is stored in the browser of the user and that can be read and written by a website to allow user and action identification. Both explicit and implicit measures have benefits and shortcomings. Explicit sign-in gives the user control over when and where the application knows who the user is. However, on the contrary, it can interrupt the user's browsing. The use of implicit techniques, such as cookies or IP address detection, allows non-disrupted identification. However, the usage is highly controversial due to privacy issues related to cookie based tracking[32].

---

[28] http://openid.net/

[29] http://www.microsoft.com/en-us/news/features/1999/10-11passport.aspx [last accessed 3/12/2012]

[30] https://twitter.com/

[31] http://facebook.com/

[32] http://donottrack.us/

To address this challenge most websites try to establish trust by complying with national privacy laws (Kobsa, 2001). Especially related to increasing trust, different studies have indicated a positive effect if the users are given the opportunity to inspect their user model (Bull, 1997; Cook & Kay, 1994; Cotter & Smyth, 2000; Czarnowski & Kay, 2000) and to opt out of logging and/or personalisation.

Privacy and user identification concerns relate to user trust in the personalisation approach (Komiak and Benbasat, 2006) and the possibility of user model scrutiny (Kay, 2006)[33]. Especially in relation to CSP model scrutiny can prove to be challenging due to the multiple source, timelines and types of information needs. This can be addressed through visualisation techniques. The discussion of information visualisation however is out of scope for this thesis. For an in-depth discussion about model visualisation strategies note Uther and Kay (2003) and Wood et al. (1997)

2.4.5. Summary

This section discussed the State of the Art in three areas related to identifying, understanding and assisting the web user in assisting information needs. The three subsections were: (1) Understanding and assisting the web user in addressing cross-site information needs, (2) modelling the user's cross-site information needs and (3) user identity, online privacy and trust. In the following the discussion in each section is summarised.

The discussion related to understanding and assisting the user in addressing information needs initially focused on what information needs are and how it originates. Based on the State of the Art discussion two key aspects can be extracted: (1) Information needs can be

[33] Further research related to the necessity of user influence on the inferred user profile is presented by Uther and Uther (2001) by adding reasons, such as access to and control over personal information, programmer accountability, correctness and validation of the model, machine predictability and aid to reflective learning. Uther and Uther (2001) as a second example discuss profile scrutiny requirements as providing overview for the user, relevance indications for the user to identify and examine relevant parts of the profile, navigational aspects, component differences showing different component types (knowledge, beliefs, etc), component values and component value confidence to show how confident the model is about the value for the component.

described as a state in which the user's knowledge underlies a certain incompleteness. (2) Users tend to address this incompleteness through a sequence or process of interactions that supports sense making. The main challenge in relation to CSP is therefore related to supporting this sense making process, which assists the user in overcoming the notion of incompleteness.

This initial discussion was followed by investigating on how to identify the user's cross-site information needs. This discussion centred on the two different approaches of implicit and explicit user data collection. Based on the notion of browsing that underlies CSP (as opposite to searching) the State of the Art discussion indicated that the usage of implicit data collection might be more appropriate. The main reason for this is (1) the continuous collection of information and (2) not requiring the user explicitly interrupt browsing to rate an item or page of interest. The main challenge identified is in relation to idle time and how the user interactions (or not interactions) can be mapped to specific needs or preferences.

Finally, State of the Art research in how to assist the user was discussed. For this research of Taylor (Taylors filters) and the Lumière Project was discussed. This discussion was focused on the challenge of balancing the assistance mechanism between *helping* and *getting the way*. In relation to a CSP this challenge triggered a discussion on what assistance can be provided, e.g. link or content-based assistance, and how an approach can identify that assistance is required. A possible indication discussed was the identification if the user is lost. An overarching challenge for CSP therefore can be describes as identifying what assistance is best to balance helping, and getting in the way, and furthermore when should this assistance be offered.

The next section discussed the modelling the user's information needs. Here different topics related to user modelling were discussed. The topics were user profile types, modelling the user across separate websites, cold start problem, modelling temporal aspects, model gaps and modelling different information needs.

In relation to user model types three use profiles were discussed (keyword based profiles / semantic network profiles and concept profiles). The main challenge identified in relation to CSP is that a user profile requires an overarching understanding of the user's browsing

space. This overarching understanding (described above as shared conceptualisation) can be modelled in all three types (which also refers to the discussion about modelling the user across separate websites). However, it should be balanced with the need to semi-automatically (requiring the user's participation) or automatically update the model. Further challenges discussed relate to the cold start problem that may occur with new users or new information needs; Modelling temporal aspects to identify how timely and relevant an identified need is; Dealing with model gaps that may occur if the user decides to disabled the approach on some website even though the information is related to the overall information needs and finally modelling separate needs and dealing with spontaneous switching between those different needs. These separate, but interconnected, user modelling challenges require a reflection in the design of a CSP approach.

Finally, user identification, online privacy and trust were discussed. Here, the main challenge lies between allowing the user to participate in the process through mechanisms such as single-sign-on or less intrusive, but privacy critical browser cookie.

The following section discussed the State of the Art related to content related to the web content dimension of Web Personalisation.

2.5. Web Content Dimension

"When it becomes evident that the elastic properties of available materials had a great deal to do with the bow, he branches off on a side trail which takes him through textbooks on elasticity and physical constants. He inserts a page of longhand analysis of his own. Thus he builds a trail of his interest through the maze of materials available to him."
(As we may think by Vannevar Bush, 1945)

2.5.1. Introduction

In the following the focus lies on discussing methods and techniques that allow an assistance mechanism to understand or categorise information within a webpage. To structure this discussion this section is divided into the two sections of *pre-structured* and *unstructured* web content (Catledge and Pitkow, 1995; Feldman and Sanger, 2006; Pol et al., 2008).[34] After this a brief summary of open and closed corpora is added. This section concludes with a summary.

2.5.2. Extracting Meaning from Pre-Structured Web Content

It can be argued that web content, by definition, underlies a pre-defined structure based on the underlying Hypertext Mark-up language (HTML). The most basic tag, that can be used for categorisation is the anchor tag[35]. This tag allows the web browser to jump to a specific section of the webpage. Assuming that the designer of the webpage has used sensible anchor names it is possible to gain an initial understanding of the underlying information. A system using html anchor tags for personalisation is ML Tutor (Smith and Blandford, 2003). ML Tutor recommends related web links based on the user's current browsing history. This

[34] Based on the web focused nature of this thesis the focus lies on semi-structured and unstructured content. It should be noted that both terms (structured and semi-structured content) are often used similar in the literature. In the context of this thesis structured content is based on a database (tabled) structure that is well well-defined (Arasu and Garcia-Molina, 2003). Semi-structured content is based on an underlying structuring standard or method to assist presentation and organisation (Abiteboul, 1997). Finally, unstructured content relates to text, images, music or videos which do not underlie any pre-set structure or semantic interpretation (Buneman et al., 1997).

[35] http://www.w3.org/TR/html4/struct/links.html

is done through a conceptual clustering algorithm introduced by Hutchinson (1994) that allows the clustering of web pages that contain similar attributes. The attributes are pre-extracted from the web corpus based on HTML anchor tags. Han and Elmasri (2004), as a second example, use the DOM (Document Model Object) structure of a HTML based webpage to create a searchable taxonomy representation to allow further analysis. In addition HTML includes a meta-tag with the attribute *name* to provide more insight about the nature of the content. [36] Examples of name attributes are *keywords*, *description*, and *author*. Within the HTML5 standard additional attributes have been introduced, such as *charset*, *content* and *name*.[37]

To allow the adding of categorisations to HTML different metadata standards have been introduced to HTML. Currently the most prominent are known as *Microformat*, *RDFa* and *Microdata* (Bizer et al., 2009).[38]

Even though HTML is continuously evolving, it can be argued, that its main purpose is not to enable the description of information, but to ensure the information is rendered by different web browsers according to the mark-up.

A standard focused more on the structure and not on the presentation of web content is *XML* (Extensible Mark-up Language). The main advantage of XML over HTML, in terms of

---

[36] http://www.w3.org/html/

[37] http://www.w3.org/html/wg/drafts/html/master/

[38] *Microformat* is used by Google and Yahoo within their indexing service since 2009. It is defined by a small set of different Microformats depending on the purpose of the information it should express. Each microformat consists of predefined attributes. One example is the microformat *vcard* with attributes, such as *fn* (first name), *org* (organisation) and *tel* (telephone number). Applications can use this information identify which HTML tag represents a vcard that related to a person. Several additional Microformats were introduced with several under development. The main limitation of Microformat is that it cannot represent any kind of data outside of the given fixed Microformats. *RDFa* provides a more flexible approach by enabling RDF (Resource Description Framework) based relationships into HTML pages. This is done by labelling content to express a specific type of information. In RDF these information types are defined as entities or items. Using the example above the entity person has the properties first name, organisation and telephone number. The important difference between microformat and RDFa is that RFDa can use any standardised vocabulary, such as DublinCore. The shortcoming of RDFa however is the potential diversity in what vocabularies are used, which can cause miss-matches disambiguaty (e.g. name can mean the full name or only the sure name). Finally, Microdata consists of five global attributes to allow relationships within the information, such as *itemid*, *itemtype* and itemscope. Furthermore any custom vocabulary can be used. One of the shortcomings of Microdata is that it only validates within HTML5.

data structuring, is that the tags, used to describe the information, are not pre-set. This allows the creation of a flexible tree like data representations. Therefore XML is useful for information, which requires a semi-structured representation, such as dictionaries, product catalogues or navigation maps. Furthermore tools can use this structure to parse and extract information from the semi-structured representation.

To ensure re-usage and inter-operability standards based on XML have been introduced. A standard based on XML is *DocBook*[39]. The main usage of DocBook is to provide a structure for content related to product manuals, dictionaries, books' or help sections. DocBook allows the capturing of a logical structure within content. For this, different structural tags are used, such as *set*, *book*, *part*, *article*, *chapter* and *appendix*. Sah and Wade (2010) propose a semi-automatic metadata extraction approach based on DocBook.

In the following more open web and unstructured approaches to extract meaning from content are discussed.

2.5.3. Extracting Meaning from Unstructured Web Content

To enable Web Personalisation on the open web, methods of extracting meaning from unstructured web content need to be investigated and discussed. Typical examples of unstructured content are among others e-mails, text documents, audio and video recordings.[40] In relation to web content typical examples are user generated content, such as blogs, forums, but also social content, which underlies high frequency updating and the use of colloquialism (Parikh and Movassate, 2009). It can be argued that extracting meaning from unstructured content is more difficult than extracting meaning from structured or semi-structured content. The main reason for this is that the content is mostly added in free text and rarely structured based on well-defined labels or relationship structures, such as taxonomies or ontologies.

To achieve this several different approaches have been introduced. One of the most studied automatic/semi-automatic approaches are known as Text Mining. The objective of text

[39] http://www.docbook.org/

[40] Within the scope of this thesis the focus lies on text based web content. However, Cross-Site Personalisation is not limited to text based web content as long as machine readable tags or annotations describing the web content are provided.

mining techniques is to identify and explore patterns of interest within the text. These patterns are either recognised over time by learning from the user usage data or directly by analysing the syntactical structure (Tan, 1999).

In the context of the web, text mining is often referred to as web content mining and can be defined as the process of discovering potentially useful and previously unknown information or knowledge from web content (Johnson and Kumar Gupta, 2012; Kosala and Blockeel, 2000). A specific area of web content mining, known as Information Extraction, enables the identification of relevant information within unstructured text. This is done by ignoring extraneous and irrelevant information (Cowie and Lehnert, 1996). The main advantage of Information Extraction is that after analysing the text only relevant parts remain (Cunningham, 2005). The main shortcoming is related to the processing time. Most Information Extraction methods rely on off-line components to produce relevant models and rules for online-processing (Cowie and Lehnert, 1996).

In recent years several tools have been introduced that assist the process of Information Extraction. The Generic Architecture for Text Engineering (GATE)[41] (Cunningham, 2002) as one example consists of three elements: A document manager, a graphical interface and a collection of reusable objects for language engineering. The unique feature of GATE is the reusable objects, which can consist of a tagger, a parser and a dictionary. However, approaches relying on external tools require considerable effort in both training the model and integrating the extracted information into websites. A similar approach is taken by Weka[42], which uses machine-learning techniques to train an initial model. This model can be used to determine the correlation between extracted terms in the model and newly added documents. In addition to offline tools semi-automatic topic tools have been introduced. Examples are OpenNLP[43] and OpenCalais[44]. The advantage of using semi-automatic information extraction tools is the usage of a simple API, which allows quick and seamless content identification. The main shortcoming is related to the training set used. A more

[41] http://gate.ac.uk/news.html

[42] http://www.cs.waikato.ac.nz/ml/weka/

[43] http://opennlp.apache.org/

[44] http://www.opencalais.com/

flexible approach recently introduced as the Apache Stanbol project[45] uses dbpedia as main database to train initial models. The flexible architecture however allows the introduction of additional datasets and languages.

2.5.4. Closed and Open Corpus

Within the research field of Adaptive Hypermedia (AH) content is mostly related to a model that represents the content and allows interaction with other models, such as a user model and/or adaptive model (Conlan, 2004). To express this interconnection between content and content model AH research often uses the terminology of closed and open corpus. Brusilovsky and Henze (2007) defines closed corpus as:

"[…] closed corpus of documents, where documents and relationships between the documents are known to the system at design time."

Open corpus is defined as:

"[…] a set of documents that is not known at design time and, moreover, can constantly change and expand"

The discussion about open and closed corpus is closely related to the discussion of structured and unstructured content above (2.5.2 and 2.5.3). Due to the design time nature of closed corpus it usually relates to a pre-defined structure. On the contrary open corpus content is mostly unstructured and not known at design-time (as described in the definitions above).

This research extends the definitions of open and closed corpus above by including the application on which the content type is hosted. Therefore this thesis relates to closed corpus (not known and defined at design-time) as hosted within **closed domains** and open corpus (known and defined at design-time) as hosted within **open domains**.

[45] http://stanbol.apache.org/

2.5.5. Summary

This section discusses and identifies challenges related to the use of web content in a CSP approach. The main focus was on studying techniques that enable the extraction of meaning from web content to enable a unified understanding of the user interest across separate websites. For this three sections were introduced. The first section discussing the extraction of meaning from web content that underlies a pre-set structure (e.g. HTML and XML). The second discussing the extraction of meaning from unstructured web content. And finally a brief discussion was added in relation to open and closed corpus.

The first section discusses different approaches using HTML and XML to extract meaning from web content. A challenge for CSP in relation to extracting meaning from semi-structured content is the identification of meaning that is added after analysing the content. Machine learning approaches are discussed in relation to this challenge; these approaches however require a good understanding of the overall content to allow the creation of a trained model. For this the methods of text mining and related tools were introduced and discussed. The main challenge identified in relation to extracting meaning from unstructured content relates to identifying possible approaches and techniques best suited for open domain use cases.

The next section discusses approaches and techniques related to the Web Application Dimension.

2.6. Web Application Dimension

"It was an impressive achievement, of course, and a human achievement […], but Deep Blue was only intelligent the way your programmable alarm clock is intelligent. Not that losing to a $10 million alarm clock made me feel any better."
(Grandmaster Garry Kasparov quoted in interview with Chess Metaphors: Artificial Intelligence and the Human Mind (2009)).

2.6.1. Introduction

Three main areas can be identified in relation to investigating the applicability of personalisation methods and approaches for CSP. The first area is related the personalisation mechanism not interfering with the user browsing freedom (non-intrusive) i.e. the personalisation should not lead to an effect in which the user is guided or steered towards content that may not be of interest (Bohnert and Zukerman, 2009). This area relates to the second area which is that a CSP approach should ensure that the website's look and feel is not aversively affected (non-invasive) in a way the website publisher may not want (Koidl et al., 2009). Both areas however can create a conundrum between the needs of the user and the needs of the website publisher. [46] For example, it can be of significant economic value for the website publisher to guide a user towards content that is promoted and may not be useful for the user. Furthermore, a website publisher may want users to spend more time within one website and not be assisted across separate websites. It therefore can be argued that a CSP approach requires addressing both areas by seeking a *state of equilibrium* in which the needs of the user and the website publisher are balanced (Equilibrium of Web Personalisation). The third and final area discussed is *interoperability*.

[46] This two-sided challenge has also been reflected in HCI research where (Berendt and Spiliopoulou, 2000) argue that Web Personalisation should be engineered to not interfere with the underlying look and feel of the website. This is because a website is not a mere collection of web pages rich in information or in product offer. It is a network of structurally or semantically interrelated nodes, usually built in a way that reflects the web designers' intuition. If the browsing experience, due to external influence, diverges from the visitor's intuition neither carefully gathered content nor sophisticated web page design guarantee successful distribution of information or goods. This argument reflects similar notions in Human Computer Interaction (HCI) research (Brusilovsky and Rizzo, 2003; Fleming and Koman, 1998; Kalbach, 2007).

This requires an investigation on what type of techniques and methods enable an assistance approach to be useful across independently hosted website.

The three overarching areas described above are illustrated in the following Figure 3:

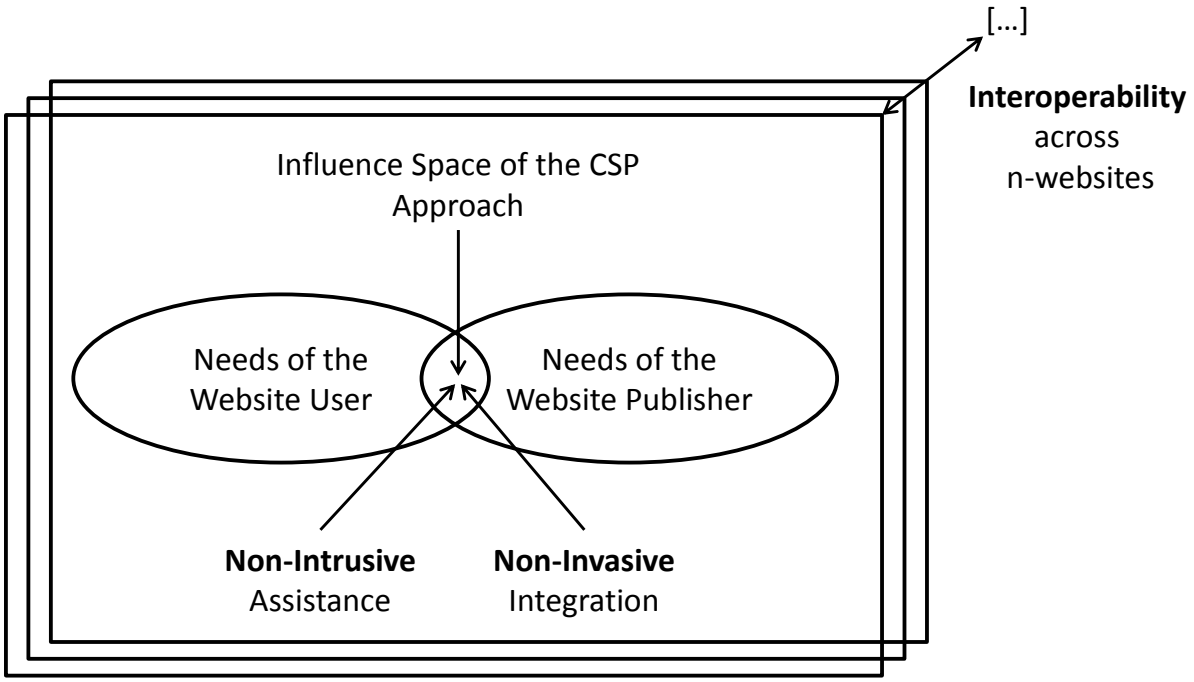


Figure 3 Overarching CSP application areas to address

This chapter is divided into sections. First, challenges in providing non-intrusive browsing assistance through personalisation are identified and discussed (2.6.2). This is followed by a State of the Art investigation on integrating personalisation into websites in a non-invasive fashion (2.6.2). And finally, a State of the Art discussion in providing interoperable personalisation techniques and methods is provided (2.6.3).

2.6.2. Non-Intrusive Web Personalisation Techniques

The use of the term non-intrusive derives from the motivation to apply personalisation in a way that a user is not hindered or distracted from their current task, yet still receives assistance if needed (Koidl et al., 2009). It can be argued that the level of intrusiveness a user seeks is based on subjective preferences. The following discussion seeks to focus on technical issues by discussing how intrusive current personalisation techniques and methods

are. This discussion includes a brief introduction in the categorisation and functionality of different techniques and approaches.[47]

Within Web Personalisation three techniques stand out. These are known as *Information Retrieval, Recommender Systems* and *Rule Engine based techniques*. All three techniques have proven successful in personalisation use cases, such as:

- Rule Engines in Adaptive Hypermedia Systems (Conlan and Wade, 2004) in section 2.6.2.1.
- Information Retrieval in personalised search (Feuz et al., 2011) in section 2.6.2.2.
- Recommender Systems for content and item recommendations (Linden et al., 2003) in section 2.6.2.3.

These techniques are not distinct, for example Content-based Recommender Systems are usually based on Information Retrieval techniques (Pazzani and Billsus, 2007). All these approaches have been extended towards personalisation since they have been introduced (Burke, 2002; Burke et al., 2011). In the following all three techniques are discussed and studied in relation to their level of intrusiveness.

2.6.2.1. Information Retrieval

Traditionally *Information Retrieval* (IR) has a strong focus on technological advancements, such as performance, scalability, precision and recall (Salton and McGill, 1986), which usually does not focus on the user's needs beyond a stated query. Recently, more user focused research has been introduced as *Personalised Information Retrieval (PIR)*. Techniques related to PIR seek to learn from the user's search history to build a user model. This model is then used to perform *query adaptation* and *result adaptation*. Query adaptation is mostly divided into query expansion and query relaxation (Xu and Croft, 1996). In a short query, query expansion can be applied to add related terms to the query without the user's explicit knowledge or participation (Mitra et al., 1998). In a long query the technique of query relaxation is used to reduce the amount of terms that may be

[47] Due to the wide application area that needs to be discussed for CSP it should be noted that this state of the art investigation focuses on the applicability of approaches and does not discuss or compare technical details such as algorithms or integration languages.

synonymous or confusing (Elbassuoni et al., 2011; Zhou et al., 2007). Experiments by Kumaran and Allan (2008) indicate that the use of query adaptation can provide performance improvements of up to 40%. Result adaptation, as a second technique used in PIR, relies on a user model to assist the re-ordering of the result list (Speretta and Gauch, 2005).

The main shortcoming of Information Retrieval is to limit the expressiveness of a query. It can be argued that a query only provides a partial representation (i.e. snapshot) of the user's overall information needs (Huberman et al., 1998). Nevertheless, for information needs in which the user is seeking a specific fact or piece of information an Information Retrieval based search engine can provide fast and accurate results.

In relation to the intrusiveness of PIR, it can be argued, that the introduced techniques are non-intrusive. This argument is based on the user not noticing the extension or relaxation of the query and therefore is not interrupted. On a higher level however it can be argued that search within a browsing task is in itself intrusive due to it requiring an interruption of the browsing process. Furthermore it can be argued that query expansion and relaxation mostly happens without the consent of the user. Even though the intentions of PIR are to assist the user it may result in the filtering of content that is relevant to the user.

2.6.2.2. Recommender Systems

Recommender Systems follow a similar *'finding a needle in the haystack'* notion as Information Retrieval. The main difference, however, is that Recommender Systems do not require an explicit query stated by the user. To recommend an item of interest Recommender Systems rely either on in-depth knowledge of the content or on explicit feedback of the user (such as rating). The literature discusses different types of Recommender Systems. The main approaches discussed are Collaborative (Linden et al., 2003; Schafer et al., 2007) and Content-based Recommender Systems (Pazzani and Billsus, 2007). In addition Hybrid Recommender Systems (Burke, 2002) and recently Social Recommender Systems (Guy and Carmel, 2011) are discussed.

The main advantage of Collaborative Recommendation Systems is that they do not require any machine-readable description of the underlying content to ensure effective use.

Collaborative Recommender Systems require explicit ratings data provided by the individuals interested in the item. This type of recommendation can be described as 'people-to-people correlation' resulting in statements such as ‘People that like this item also like the following items’ (Schafer et al., 2007, 2001).

Content-based Recommender Systems do not require user ratings. They are based on a query that is automatically created based on the user information stored in the system (Pazzani and Billsus, 2007). Similar to result adaptation in Personalised Information Retrieval, the resulting list is used to inform the Recommendation System about the relevance of items in relation to the user’s preferences and needs.

Recently variations of Recommender Systems have been introduced, such as Hybrid Recommender Systems using both rating and information retrieval data (Burke, 2002), as well as Social Recommender Systems, which only use ratings from friends and peers within the social network of the user (Guy and Carmel, 2011).

- Recommender Systems have several shortcomings. The main shortcomings are known as *New User* (also known as *Cold Start Problem*), *New Content* problem and *Portfolio Effect*.
- The *New User problem* relates to missing information about a new user (e.g. no rating or/and no click data) resulting in a higher probability for misinformed recommendations (Burke et al., 2011).
- The *New Content problem* relates more to Content-based Recommender Systems and can occur if new content is added, but not incorporated into the underlying model/system that calculates recommendations. This problem is apparent in large and dynamic content base, such as news websites, which underlie rapid content updates (Herlocker et al., 2004).
- Finally, the *Portfolio Effect* relates to Recommendation Systems recommending items or content already consumed/purchased and therefore are not relevant anymore. This shortcoming has been addressed by allowing the user to actively participate in the creation and updating of the user’s profile (Schafer et al., 2002). Other solutions have been introduced for example by adding demographical data, such as age and gender (Webb et al., 2001), specifically to overcome the *New User*

problem. Some strategies relate to implicit techniques, such as Web Usage Mining or Conceptual User Tracking (Oberle et al., 2003). Both techniques rely on analysing log data for unique evidence such as IP address, device preferences and navigational behaviour (Mobasher et al., 2000).

The level of intrusiveness introduced by Recommender Systems varies and depends on how much the user is required to participate. Therefore it can be argued that Content-based Recommender Systems are less intrusive than Collaborative Recommender Systems, which mostly rely on explicit rating. Both approaches have their use especially when trying to find specific items/facts, however in information needs that are not specific and possibly span over different sources requires the user to cognitively connect the partial information. This can result in browsing patterns that consists of back tracks leaving the browsing experience shallow (Herder and Juvina, 2004).[48] A further concern in relation to the intrusiveness of the Recommendation Systems and Information Retrieval has been identified as 'over filtering' or 'over personalisation' and is also referred to as the *filter bubble* (O'Callaghan, 2011; Pariser, 2011). This shortcoming can lead to the user receiving more of the same without any diversity or serendipity in either the search results or the served recommendations.

2.6.2.3. Rule-based Systems

Rule based systems, as a third approach, mostly rely on closed and pre-set environments with allowing the applicability of static conditions. Rule Engine based techniques are mostly used in subject specific application where the content does not require frequent updating. One example of an application area is *Adaptive Hypermedia Systems (AHS)*, traditionally focuses on education. Here, the level of intrusiveness is high, based on the necessity to guide the user to content that is appropriate in relation to the user's level of knowledge

[48] This discussion relates to the notion of *teleporting* and *orienteering* as introduced by Jansen et al. (2008) is used. Teleporting refers to navigational support which teleports the user to a specific page or item. Examples of this are clicking on a search result within a search engine or clicking on a recommended item within a retail website. The main shortcoming of teleporting is that in exploratory and broader information needs the user can only address one aspect of this need. The user is teleported directly to the page for which the information retrieval algorithm has calculated the highest relevancy. The same applies for recommender systems. The paradigm here is to recommend content that has been identified as closest to the current page or item of interest. Both approaches use teleporting although recommender systems do not rely on a set of keywords provided by the user (Schafer et al., 2007).

(Conlan and Wade, 2004). A second example in the use of rule engines is gaming. The conditional rule sets of the rule engine can provide alternative game paths depending on user's level of knowledge and/or experience (Koidl et al., 2010). Even though games tend to simulate an open domain by allowing the user to freely move and explore, gaming environments usually do not underlie any sudden changes throughout their life cycle.[49]

The main advantage of rule engines is that they are fast and reliable. However, creating a rule set is time consuming. Furthermore, it requires a strong interconnection to the underlying logic of the application. Changes or updates in the application may result in complex rule updates. This makes rule engines difficult to apply in open domains that undergo rapid changes.

Rule engine approaches can vary in intrusiveness based on the application area. This is due to Rule Engines only assisting the decision process based on the use of a conditional rule set. Based on this inherent conditionality rule engines enforce a binary decision that in some environments enable more intrusive decisions on behalf of the user. For example, in a learning environment the intrusiveness imposed is usually high due to actively guiding the user to content that is appropriate for the user's level of knowledge. In gaming environments the challenge of non-intrusiveness becomes complex. Here a fine balance between guiding the user and not interfering with the game flow should be considered (Koidl et al., 2010).

The following section discusses challenges in integrating Web Personalisation techniques in a non-invasive manner.

2.6.3. Non-Invasive Integration of Web Personalisation Techniques

As introduced in chapter one of this thesis, this thesis argues that personalisation on the web, especially in relation to browsing across separate websites, needs to address, and if possible balance, the needs and preferences of the user and the needs and preferences of the website publisher. The former is discussed above by discussing the notion of non-intrusiveness. The latter is discussed in this section. The main focus herein lies in

[49] Due to their speed, rule engines are applied in various time sensitive applications, such as financial transactions and health care. These applications are out of scope and not related to Web Personalisation.

investigating the State of the Art in integrating Web Personalisation techniques into existing websites. It is argued that this integration should be introduced as minimal as possible. Minimal, in this context, relates to the invasiveness of the integration. Moreover, this thesis argues that a non-invasive integration does not aversively interfere with the look and feel of the websites and the control the website publisher.

To discuss the challenge of non-invasive integration of Web Personalisation techniques into a website following two application areas are studies:

1. Non-Invasive integration of Web Personalisation into websites at design-time (2.6.3.1).
2. Non-Invasive integration of Web Personalisation into websites at run-time (2.6.3.2).

The overall goal of this discussion is to identify the level of invasiveness necessary to allow a consistent assistance of the user across individually hosted websites.

2.6.3.1. Non-Invasive integration of Web Personalisation at design-time

The integration of Web Personalisation at design-time is complex. This is due to the different standards, specifications and technologies that are used to design a website. This diversity makes the integration of further techniques difficult.

A research and application field that addresses the shortcoming of integrating techniques and methods to existing websites is known as Web Engineering. This field introduces a structured modularisation of the website development process (Schwabe et al., 2001). It relies on modeling the overall website based on software engineering principals. It allows a more structured design process ensuring a more rapid, structured and re-usable development of web applications. For this several web engineering design methods were introduced. The main design methods introduced, such as OOHDM (Rossi and Schwabe, 2008), UWE (Koch et al., 2008) and OOWS (Pastor et al., 2006). In addition data modeling approaches, such as WebML (Ceri et al., 2000) and WebRatio (Acerbis et al., 2007) were introduced as model-driven engineering environments.

The advantage of Web Engineering is that it relies on strict design methodologies. For example OOHDM (Object Oriented Hypermedia Design Method) has the phases:

requirements gathering, conceptual design, navigational design, abstract interface design and implementation. The result is UML like diagrams and re-usable design pattern. Most Web Engineering approaches include functionalities that assist the user in finding and exploring information. This includes techniques discussed above, such as Information Retrieval and Recommendation Systems (Ginige and Murugesan, 2001).

The main shortcoming of Web Engineering is lack in flexibility in extending the websites functionality beyond what has been stated in the websites setup design. To overcome this shortcoming Ginzburg (2009) introduces different role based views into the presentation layer of the web application by enriching the abstract data layer of OOHDM. This enables XML based representation exchange between the application and interfacing service, which is required to ensure third-party integration in relation to Web Personalisation. The approach relies on JSP pages being written in well-formed eXtensible Mark-up Language[50] (XML) to allow eXtensible Stylesheet Language[51] (XSL) transformations injecting templates that indicate where role based changes apply.

It can be stated that the level of invasiveness to extend an already deployed website, based on Web Engineering principles, is high. To introduce any additional functionality the underlying website needs to be re-designed and re-deploy. However, this limitation is balanced by the fact that a website that has been designed and implemented based on Web Engineering principles underlies strict design and documentation guidelines can reduce cost and time.

2.6.3.2. Non-Invasive integration of Web Personalisation at run-time

In contrast to non-invasive integration at design time, such as based on Web Engineering principles the integration of personalisation methods and techniques at run-time is even more challenging. The main reason is the missing understanding of the interdependencies within a website that limit flexible extensibility. A promising approach is the use of website Frameworks. Such frameworks are often referred to as Web Information Systems (WIS) and allow the implementation of an entire website as out-of-the-box deployment (Avison and

[50] http://www.w3.org/XML/

[51] http://www.w3.org/Style/XSL/

Fitzgerald, 2003; Isakowitz et al., 1998). In certain environments this approach enables non-invasive extensibility during run-time.

One class of WIS has become known as Web-based Content Management Systems (WCMS)[52]. It allows a simple implementation of a website ranging from blogs to more complex e-commerce websites. In addition to open source WCMS commercial Content Management Systems (CMS) were introduced. The investigation of the extensibility requires an extensive analysis of current approaches and trends in WCMS. For this a State of the Art survey in WCMS was conducted (appendix on page 226). The outcome of the initial survey indicated three main benefits of using open-source WCMS in relation to extension during run-time towards applying CSP. First, the simple extensibility and control of the added mechanism/s by the website publisher through a pluggable framework is constantly maturing due to an active participation by the developing community. Second, WCMS have moved into a mature development state with increased usage on enterprise level and growing adoption rates in all areas of the web.[53] Finally, WCMS implement an extendable framework, which allows external modules to influence different level of the functional layer of the website without the need of re-design or re-deployment. The main shortcomings of WCMS are a lack of strict and coherent coding guidelines and documentations that is balanced due to the active support of the developing community.

By providing an extendable core implementation that can plug different modules depending on the application area, WCMS provide a framework that caters for many different types of application areas. This high-level of flexibility and abstraction though comes with the cost that developers need to have good knowledge of the core elements in order to extend them. For example, an important feature of Drupal[54], a prominent WCMS, is to allow extensibility is hook-able architecture. This is a key feature that allows external plug-ins to flexibly 'hook' into the calling stack of the platform and in different levels of the overall architecture. *MediaWiki*[55], as a second example, provides an architecture to add useful

[52] Currently more than 10% of the top 10.000 traffic sites on the web are implemented as Web-based Content Management System according to http://trends.builtwith.com/cms sourced on the 11/04/2013.

[53] http://trends.builtwith.com/cms

[54] http://drupal.org/

[55] http://www.mediawiki.org/wiki/MediaWiki

functionality to the implementation via the *extensions* feature. These are based on simple scripts adding different plug-ins like Adobe Flash, Video streaming, RSS feeds, ratings and API based third party accessibility. Recently more enhanced extensions were introduced adding semantic operations like adding semantic structure and relationships to different articles within the content of MediaWiki.[56]

It can be stated that the integration of Web Personalisation techniques into WCMS is non-invasive due to the extensible and pluggable design of WCMS. However, this only applies if the added personalisation techniques is limited to certain areas e.g. a widget. This allows the website publisher to control the extension (if necessary disable) and also decide how much the widget should influence the overall websites look and feel. Furthermore it should be noted, that it is possible to design and enable more invasive personalisation features at run-time. Example of this can be deep link recommendations that change the presentation and delivery of content within a website. This, however, would have to be balanced with the needs and requirements of the website publisher.

2.6.4. Interoperability of Web Personalisation Approaches

Interoperability of approaches and mechanisms that assist users in addressing needs and preferences have been introduced by several research groups (Brusilovsky et al., 2005b; Carmagnola and Cena, 2006; Carmagnola et al., 2011). The main focus in relation to interoperability herein lies in allowing the user, content and adaptive model to be used in different applications and approaches. For CSP interoperability requires the discussion of seamless integration across the different websites the user browses (non-invasive). Furthermore, this interoperability needs to enable a seamless experience for the user (non-intrusive). Seamless interoperability should include interoperability across devices and independent of the websites technology stack.

This section discusses the State of the Art in different approaches related to the interoperability of Web Personalisation techniques and methods. Due to scope interoperability approaches for user models is not specifically discussed. For an overview of interoperability in relation to User Modelling view (Carmagnola et al., 2011).

[56] A complete list of official MediaWiki extensions can be found under http://www.mediawiki.org/wiki/Extension_Matrix

In the following interoperability in the application areas of Adaptive Hypermedia Systems (2.6.4.1), Scout approaches (2.6.4.2), Intermediary approaches (2.6.4.3) and browser-based plug-ins (2.6.4.4) are discussed.

2.6.4.1. Adaptive Hypermedia Systems

The focus of Adaptive Hypermedia System (AHS) is to overcome the 'one size fits all' paradigm by enabling a system to adapt to the individual needs and preferences of the users (Brusilovsky, 1996). Traditionally the main focus herein lies in assisting different users within one application and not across different applications.

According to Jameson (2009) adaptive systems usually focus on one of the following five use cases: a) Helping users to find information; b) Tailoring information presentation to the user; c) recommendation products; d) support collaboration and e) support Learning. Knutov (2009) argues that in order to design a system that addresses the needs and preferences of users within one of these use cases the following design questions need to be stated: *What can we adapt?; What can we adapt to?; Why do we need adaptation?; Where can we apply adaptation?; When can we apply adaptation?; How do we adapt?* To enable effective adaptivity different models are required. Typical models are content or domain models; user models or user profiles; strategy, learning or narrative models. The success of an AHS is based on guiding a user towards relevant content that would have been either missed on not found as effective of efficient Kay (2006). Within AHS this guidance is either based on *adaptive presentation* and *adaptive navigation support* (Brusilovsky, 2001, 1996).

AHS have clearly proven to be successful in specific subject domains and use cases, such as e-Learning applications (Brusilovsky, 2004; Conlan and Wade, 2004) and web-based navigation support (Brusilovsky et al., 1996; Henze and Herrlich, 2004). Despite this success the main shortcoming of AHS is limited to a specific, and mostly isolated, implementation. To overcome this shortcoming the interoperability of the different models within an AHS has been investigated.

In relation to content model interoperability Open Adaptive Hypermedia Systems (OAHS) have been introduced and defined as: *"An open adaptive hypermedia system (OAHS) is an*

adaptive hypermedia system which operates on an open corpus of documents" (Henze and Nejdl, 2006). This definition is extended by Brusilovsky and Henze (2007): *"[...] an open corpus adaptive hypermedia system is an adaptive hypermedia system, which operates on an open corpus of documents, e.g. a set of documents that is not known at design time and, moreover, can constantly change and expand."*

Different research groups have addressed the Open-Corpus Problem by developing more flexible architectures. A subset of the approaches discussed in the literature is KnowledgeTree (Brusilovsky, 2004), ADAPT^2 (Brusilovsky et al., 2005b), The Personal Reader (Henze and Herrlich, 2004), APeLS (Conlan et al., 2003) and AHA! (Bra et al., 2003).

KnowledgeTree, as one example, addresses interoperability by implementing a separation of concerns within the AHS by using a community of distributed servers. However, the focus of KnowledgeTree is not to provide more effective and flexible adaptive features, but to allow easier integration and re-use of content in AHS (Brusilovsky, 2004). Four different server types are introduced: *learning portal servers*, *activity servers*, *value-added service servers* and *student modelling servers*. The novelty of this approach lies in the separation of concerns to allow an easier re-use and integration of content. Extensions to KnowledgeTree, such as Adapt^2 (Brusilovsky et al., 2005b) extend this initial framework by adding an ontologically driven approach (Brusilovsky et al., 2005b).

APeLS (Conlan et al., 2003), as another example provides a distributed service based architecture and sophisticated adaptivity mechanisms. The individual services in APeLS are (a) the Adaptive Hypermedia Service providing the content and (b) the Learning Environment. The latter is used to track the learner, based on the tutor's guidance in the form of learner profiles, assessment information and pedagogical constraints (Conlan, 2004). Similar to KnowledgeTree and AHA!, APeLS relies on closed corpus, which should be integrated at design time. This makes the interoperability across multiple independent services difficult.

Another example, following the notion of separate concerns, is the *Personal Learning Assistant* (PLA) introduced by Dolog et al. (2004), which follows a service based approach similar to APeLS (Dagger et al., 2007). PLA is a service-oriented approach to personalised

learning portals by connecting different services based on the requirement of the learner. The individual services are divided into three sections, the *Personal Assistant* and user interaction, personalisation services and supporting services. To ensure communication between the services WSDL and SOAP are used. More recently personalised learning portals have been introduced that focus on more advance service composition and workflow management (Staikopoulos et al., 2012). Similar to APeLS and KnowledgeTree PLA does not explicitly address the interoperability of personalisation techniques and methods, but more on extending a portalised approach and enabling better import and export of content.

AHA! (Bra et al., 2006b), as a final example, places stronger emphasis on the adaptivity mechanism although compared with KnowledgeTree and APeLS provides a less flexible architecture. The adaptive features of AHA! are based on the Adaptive Hypermedia reference model AHAM (Bra et al., 2006a). However even though AHA! provides a rich adaptive experience, the interoperability of concepts and content is not specifically addressed. Nevertheless, the usage of a generalised reference model, such as AHAM, in the design of AHS provide a promising approach to easier decouple the adaptive logic from the content; thus allowing easier re-use and interoperability of content. AHAM addresses the challenge of facilitating more open and flexible personalisation and has been introduced by De Bra et al. (1999) and Bailey et al. (2007). The model was later successfully applied to AHA! (Bra et al., 2006a) illustrates the possibility of using an underlying design framework. In addition to AHAM (Bailey et al., 2007) the Fundamental Open Hypermedia Model (FOHM) was introduced. In comparison to AHAM the FOHM not only defines the elements of an AHS and their interconnectedness, but also how adaptive techniques can be combined to allow more simple and flexible design of AHS.

A promising approach to interoperability has been introduced by Yudelson (2010) as serviced-based approached introduced as *Adaptive Feature Service (AFS)*. Here, the adaptive functionality is outsourced to a separate service. The goal is to provide a service that allows for a more interoperable personalisation across different platforms. To facilitate the communication between the AFS and the applications a communication protocol is introduced. Therefore the main limitation of this approach is that the application needs to adhere to the predefined protocol. It is also not clear how much effort a system designer

needs to place in integrating both the protocol and adaptive features. It remains a good example of AHS moving towards a more flexible separation of concerns.[57]

Based on the discussion above it can be concluded that most approaches in AHS remain domain focused with most considerations related to the interoperability of the content model but not on assisting the user across separate systems. Limited consideration is placed on interoperability of the adaptive logic within an adaptive system, such as the underlying rule engine.

In the following scout approaches are discussed. These approaches are similar to AHS although with a stronger focus on web based systems.

2.6.4.2.Scout approaches

A scout system relies on approaches that allow the scouting of navigational paths the user has not yet taken. *ScentTrail*, for example, a system introduced by Olston and Chi (2003) implements the theory of *information scent* by Chi et al. (2001), *information foraging*[58] by Pirolli and Card (1999) and insights on *orienteering behaviour* (Fleming and Koman, 1998). Information Scent can be defined as the imperfect, subjective perception of the value, cost, or access path of information sources obtained from browsing cues (Olston and Chi, 2003). The basic idea is that while the user is engaged in information gathering activities, a *scout*[59] looks ahead. It allows the assessment of page relevance and based on how high this relevance is, the user is offered navigation suggestions. These navigation suggestions are

[57] Facebook, as a second example, has applied an approach named *Instant Personalisation*[57]. The idea of this approach is to use the social graph of the user to infer social recommendations on different website. For this the user needs to authenticate on the website with his or her Facebook credentials. This information then allows the website to indicate which links are *liked* by others within the user's social graph. The main shortcoming of this approach is the missing understanding if the link was liked based on a momentary notion or based on actually liking the item of interest. Furthermore the user has no control over the social recommendation. However, it should be noted that instant personalisation presented by Facebook is similar to the Adaptive Feature Service discussed above. It enabled interoperability in which Facebook serves as a user model service. The lack of control and transparency question its applicability beyond simple like based recommendations.

[58] Information foraging is strongly related to the theory introduced as Berry picking by Bates (2002). Both focus on interleaving structured behaviour (searching) with opportunistic and unstructured behaviour (browsing).

[59] The selection of the term scout was intentional to disseminate from the term 'agent' within Artificial Intelligent (AI) research. Agents within AI underlie a specific categorisation, such as reflex and utility agent. This categorisation does not apply here.

based on an enhanced information scent. To understand the user's initial need ScentTrail uses the user's search query. The relevance of each page in an information patch is calculated based on conventional information retrieval techniques, such as TF-IDF. *Spreading activation* (Collins and Loftus, 1975) is used to calculate the page relevance backwards along links; each page's information scent is recursively defined as the sum of its own relevancy and the relevance values of its children. This calculation also includes a decay parameter. The calculated information scent is used to highlight the links according to their informational scent.

Similar research was conducted by Lieberman (1995) with the Letizia system and WUFIS (Chi et al., 2001). In both cases the same notion of simulating or anticipating the user navigation was studied. Compared with SentTrail, Letizia takes the user's reading idle time into account. In terms of the chosen annotation techniques front size is used by ScentTrail and Letizia.

In relation to interoperability scouting approaches were not primarily developed to enable integration in different systems and/or provide a seamless browsing experience across different website. However, the core notion introduced by scouting systems, which is to enable the assistance of non-intrusive browsing, is promising. Furthermore it can be noted that the approaches applied, such as an IR based result list, can be used across websites if integrated correctly.

2.6.4.3. Intermediary Approaches (Portal approaches)

Different intermediary approaches to personalisation have been introduced, in which the user logs into a separate platform/portal in order to receive a personalised experience. The main notion is that the user can address information needs in one portal/website and not across the web on different websites. For this these systems usually consume the content of other platforms and reorganize the presentation. This allows for rich personalisation presentation across different content types (such as video and text) and also across different languages (Steichen et al., 2009). However, it mostly relies on a single site approach that may be limited in the amount of content provided (due to copyright and content harvesting restrictions). Related to interoperability intermediate approaches focus on content import to allow content aggregation.

2.6.4.4. Browser Extensions/Plug-ins

In addition to intermediary approaches *browser extensions* in the form of plug-ins have been introduces. These add personalisations into the presentation layer of the browser. One example is the 'Smarter Wikipedia' (now called Fastestfox) Firefox browser plug-in, which adds a 'related articles' box to MediaWiki's[60] Wikipedia implementation and other website such as Amazon.[61] Other examples are *StumbleUpon*[62] and *Google Related*[63]. Both add a menu bar to the lower end of the browser in which related articles are suggested. The personalised suggestion however is only based on the content of the current page, not on previous interactions or information needs that are of higher order or more abstract. Therefore similar as search engines current plugins only provide a shallow and isolated personalisation experience.

The main advantages of a client (browser) side integration of CSP are: (1) The user's information resides in one central point and (2) the cross-site-personalisation mechanism does not have to be installed on the website for Cross-Site Personalisation to function and therefore covers all websites the user browses. The main shortcomings are: (1) The user is tied to the browser(s)/device(s) on which the plugin has been installed, (2) a browser extension cannot provide the same depth of influence a mechanism has that is implemented within the websites underlying application; (3) The introduced personalisation may seem alien to the look and feel of the website potentially confusing or distracting the user.

The main functionality of a web browser is to render the web content according to the underlying mark-up. Recently various extensions have been introduced to the basic mark-up. In the context of manipulating the presentation of a webpage various scripting types have been introduced. Most notably these are *JavaScript* that is embedded into the web pages mark-up and allows the execution of a library with extended functionalities. The main focus of manipulations lies on the DOM structure of the document. With JavaScript the

---

[60] http://www.mediawiki.org/wiki/MediaWiki

[61] http://smarterfox.com/news/

[62] http://www.stumbleupon.com/

[63] http://www.google.com/related/ (This project has been discontinued)

underlying DOM structure of a webpage can be altered. This in effect allows the injection of visual elements that underlie conditional statements. Client-side *tuning* of websites (Díaz et al., 2008) based on script injection can be simplified by the use of tools. The most prominent tool is GreaseMonkey[64]. Scripts written in GreaseMonkey can be used to manipulate the DOM structure of a webpage based on the context of the browser. One example is provided by Firmenich et al. (2010) is to use client side script so introduce *concern sensitive browsing*. The concept similar to CSP seeks to assist the user on webpage based on the user's current concern.[65]

Even though concern sensitive adaptation is a promising approach, its implementation appears complicated based on the necessity to write script that can be applied for different concerns. A similar shortcoming can be identified in rule-based approaches.

2.6.5. Summary

This section discussed three areas that this thesis argues are important to allow CSP to assist the user in addressing cross-site information needs. The three areas discussed are non-intrusive Web Personalisation techniques, non-invasive integration of Web Personalisation techniques and interoperability of Web Personalisation techniques.

The discussion of non-intrusive Web Personalisation techniques was focused on the three techniques of Information Retrieval, Recommender Systems and Rule Engine based approaches.

In relation to Information Retrieval two main personalisation techniques were discussed. One focused on query extension and relaxation. The second, focused on result list adaptation, allowing a resorting of the result list. Both approaches can be stated as non-intrusive due to the user not requiring to explicitly assisting the different actions. Recommender Systems, as second technique discussed, focused on relating existing

[64] http://www.greasespot.net/

[65] One example discussed is the interconnection of concerns between a mapping application and an encyclopaedia (in this case Google Maps and Wikipedia). Because the user is browsing out of Google maps and into Wikipedia it can be concluded that the concern relates to local.

information to allow a system to build a correlation between explicit feedback (Collaborative Recommender systems), terms/tags extracted from the content (Content-based Recommender Systems), as well as Hybrid and Social Recommender Systems. In terms of intrusiveness the two main classes (Collaborative and Content-based Recommender Systems) vary with the former being more intrusive due to interrupting the user to gather required feedback and the latter less intrusive due to it working in the background, based on click and reading behavior. In addition, shortcomings of Recommender Systems were discussed. These were the New User problem, the New Content problem, the Portfolio Effect and over-filtering. Finally, Rule Engine based approaches were discussed. Here the level of intrusiveness varies depending on the application itself. Rule Engine based approaches are usually applied in closed domains relating to a specific rule set, therefore a rule-based approach may be perceived as non-intrusive due to it allowing fast and seamless adaptivity.

The next area discussed was related to non-invasive integration of Web Personalisation in websites. Non-invasive integration is discussed as an approach in which Web Personalisation is applied to a website by not aversively affecting the websites look, feel and control of the website publisher. Two different integration areas were discussed: (1) Non-Invasive integration of Web Personalisation into websites at design-time and (2) Non-Invasive integration of Web Personalisation at run-time.

The State of the Art discussion of Web Personalisation techniques at design time was focused on Web Engineering due to it introducing structure to the development process. In relation to non-invasive integration of Web Personalisation, the level of invasiveness was concluded as high due to Web Engineering not specifically focused on providing a platform that enables non-intrusive extension of websites. The second integration type discussed was integrating Web Personalisation at run-time. Here, the outcome of an initial State of the Art survey in Web-based Content Management was discussed. This class of Web Information Systems allows the extension of websites at run-time in a non-invasive manner.

Finally, the interoperability of Web Personalisation techniques was discussed. Here the focus was on Adaptive Hypermedia Systems, scout approaches, intermediary approaches and browser-based plugins. The overarching challenges identified in all three discussed

areas (integration during design-time, run-time and interoperability) related to CSP are to ensure Web Personalisation techniques can be introduced in a non-invasive manner. This includes minimising the impact the approach has on the underlying website in terms of look, feel and control for the website publisher. Furthermore non-invasive integration requires a balanced approach in terms of time and cost. Ensuring the wide distribution of Web Personalisation on the web requires a simple and non-invasive integration. From a user perspective this integration should include interoperability across websites that hosted separately. This balance, between the website publisher needs (non-invasive integration) and the website user's needs (non-intrusive and interoperability), is a challenge that should be addressed by the design and implementation of a CSP approach. In addition to aspects related to non-invasive integration of CSP to enable interoperability different approaches were discussed that address device and technology stack independence. Especially the latter is important to ensure updates to the technology stack of the website do not aversively affect the appropriateness of the CSP approach. A further overarching challenge is to ensure that any personalisation applied, within the different websites the user browses, is accurate and timely. This temporal aspect has been discussed in relation modeling the user's information needs. A main overarching challenge therefore within the Web Application Dimension is to ensure the personalisation approaches applied ensure fast and accurate personalisations. This is especially relevant in browsing spaces that underlie constant changes in their content base.

2.7. Conclusions

The State of the Art review above focused on the three previously introduced dimensions (web user/web content/web application). Within each section State of the Art approaches and methods were discussed. The goal of each discussion was the identification of challenges related to assisting users in addressing information needs that span independently hosted websites. Each section was concluded with a summary overview of the identified challenges. In the following chapter the influences of the State of the Art are extended towards specific design requirements. These requirements are subsequently used throughout the design, implementation and evaluation of this thesis.

This State of the Art review identified several challenges within Web Personalisation. The main challenges identified resulted in a set of requirements that is introduced and discussed at the beginning of the following chapter (3.1).

3. Design

This chapter proposes the design to Cross-Site Personalisation (CSP). It is based on the outcome of the State of the Art review in chapter two of this thesis and subsequently applied in two separate, but interconnected use cases. The first related to CSP in open domains the second in closed domains.

This chapter is structured as followed: First, the influences of the State of the Art review are discussed (3.1.). The outcome of this discussion is a set of design requirements (3.1.4). After this a high-level design to CSP is proposed (3.2). This design is subsequently extended in two use case based designs. The first focused on closed domains (3.3), the second on open domains (3.4). Both use cases are discussed in separate sections. Finally, a conclusion is provided (3.5).

3.1. Influences from the State of the Art

The State of the Art analysis conducted in chapter two of this thesis addressed various aspects related to CSP. The main goal of this section is to provide an overview of the influences from the State of the Art review resulting in design requirements. The requirements are introduced and addressed with varying degrees of importance and impact.

The following three sections are structured based on the State of the Art chapter of this thesis. First, requirements related to the Web User Dimension (3.1.1) are discussed. After this requirements related to the Web Content Dimensions (3.1.2) and finally requirements related to the Web Application Dimension (3.1.3).

Each requirement is introduced with an identifier. This identifier is then used in a table (Table 1 on page 64) presented in the summary of this section (3.1.4). The table provides an overview of the identified requirements and is consequently used and extended throughout this thesis.

3.1.1. Web User Dimension Requirements

The State of the Art discussion related to the Web User Dimension (2.4) was split into three areas. The first related to understanding and assisting web users in addressing information

needs (2.4.2). The main challenge identified was related to ensuring CSP maintains the right balance between *helping* and *getting in the way*. The following design requirement was identified:

- A CSP approach should introduce assistance that ensures the user is not hindered in freely browsing. [WUD-1]

The second area discussed was related to the modelling of the user information needs (2.4.3). The two challenges identified were the need for a shared conceptualisation of the user's browsing space and challenges in modelling the users information needs across independently hosted websites (among others the new user problem, new interest problem and modelling temporal aspects to identify the freshness of the need). The following design requirements were identified:

- A CSP approach should have a unified understanding of the user's browsing space by creating a shared conceptualisation of this understanding. [WUD-2]
- A CSP approach should apply a user modelling technique that does not interrupt the user's browsing. It should therefore be implicit by not requiring the user to explicitly participate in the identification, collection and management of information needs. [WUD-3]

The final subject area discussed in the Web User Dimension was related to user identification, online privacy and trust (2.4.4). Here the high-level challenge identified related to ensuring the user's privacy is honoured by ensuring websites can't access the user's cross-site information needs. Furthermore, privacy concerns relate to the user's trust were identified in studied in relation to user model scrutiny (Kay, 2006; Komiak and Benbasat, 2006)[66]. Following requirements were identified:

[66] Further research related to the necessity of user influence on the inferred user profile is presented by Uther and Uther (2001) by adding reasons, such as access to and control over personal information, programmer accountability, correctness and validation of the model, machine predictability and aid to reflective learning.

- A CSP approach should honour the user's privacy needs by ensuring the different websites the user is browsing do not receive any information about the user cross-site information needs. [WUD-4]

3.1.2. Web Content Dimension Requirements

Within the Web Content Dimension (2.5) two areas where discussed. The first related to web content that underlies a pre-set structure (2.5.2) and the second related to web content that is unstructured (2.5.3). Following requirements were identified:

- A CSP approach should be able to utilise existing content models to create a shared conceptualisation across different websites. [WCD-1]
- A CSP approach should be able to use additional services and tools for term identification if necessary. [WCD-2]
- Depending on the domain (open or closed) the CSP approach should be able to identify terms of content that is added or updated within a CSP enabled website. [WCD-3]

3.1.3. Web Application Dimension Requirements

The Web Application Dimension (2.6) was split into three areas: Non-intrusive application of Web Personalisation techniques (2.6.2), Non-invasive application of Web Personalisation techniques (2.6.3) and finally interoperability of Web Personalisation techniques (2.6.4). The overall goal of the discussion was to identify challenges and requirements of Web Personalisation techniques related to CSP.

In relation to non-intrusive integration of Web Personalisation (2.6.2), the main focus was on analysing current personalisations techniques towards their level of intrusiveness. Following requirement was identified:

- A CSP approach should provide non-intrusive assistance techniques to ensure the user is not hindered in freely browsing. [WAD-1]

In relation to non-invasive integration of Web Personalisation to existing websites (2.6.3) the focus of the State of the Art discussion was on studying in how far Web Personalisation

techniques can be applied without affecting the look and feel of the website. Based on this following design requirements for CSP were identified:

- A CSP approach should affect the website as minimally as possible. [WAD-2]
- A CSP approach should not negatively influence core areas of the website (loading time etc.). [WAD-3]

In relation to interoperability different Web Personalisation techniques (2.6.4) were discussed. The discussion focused on identifying interoperability challenges. Following CSP design requirements were identified:

- A CSP approach should be based on a design that allows simple integration and interfacing with existing websites. [WAD-4]
- A CSP approach should ensure that the communication between the CSP approach and the independent websites is flexible and does not depend on the websites technology stack. [WAD-5]
- A CSP approach should ensure device and browser independence. [WAD-6]

3.1.4. Summary

This section discussed the influences of the State of the Art resulting in a set of design requirements for a CSP approach.

Following table (Table 1) provides an overview of all design requirements identified above. This table is consequently used and extended in the design and implementation of two use cases investigating CSP in closed and open domains. The table can be found in the summary of each section. Furthermore, the table is applied in the conclusion of evaluation chapter (5.7) discussed how much the individual requirements were assessed and met.

Web User Dimension (3.1.1)	
A CSP approach should propose assistance that ensures the user is not hindered in their current browsing.	[WUD-1]
A CSP approach should have a unified understanding of the user's browsing space by creating a shared conceptualisation.	[WUD-2]
A CSP approach should apply a user modelling technique that does not interrupt	[WUD-3]

the user's browsing. It should therefore be implicit by not requiring the user to explicitly participate in the identification, collection and management of information needs.	
A CSP approach should honour the user's privacy needs by ensuring the different websites the user is browsing do not receive any information about the user cross-site information needs.	[WUD-4]
Web Content Dimension (3.1.2)	
A CSP approach should be able to utilise existing content models to create a shared conceptualisation across different websites.	[WCD-1]
A CSP approach should be able to use additional services and tools for term identification if necessary.	[WCD-2]
Depending on the domain (open or closed) the CSP approach should be able to identify terms of content that is added or updated to website as soon as possible.	[WCD-3]
Web Application Dimension (3.1.3)	
A CSP approach should provide non-intrusive assistance techniques to ensure the user is not hindered in freely browsing.	[WAD-1]
A CSP approach should affect the website as minimally as possible.	[WAD-2]
A CSP approach should not negatively influence core areas of the website (loading time etc.).	[WAD-3]
A CSP approach should be based on a design that allows simple integration and interfacing with existing websites.	[WAD-4]
A CSP approach should ensure that the communication between the CSP approach and the independent websites is flexible and does not depend on the websites technology stack.	[WAD-5]
A CSP approach should ensure device and browser independence.	[WAD-6]

Table 1 High-Level Design Requirements

The following section introduces the high-level design of the CSP approach. The table above is extended in the summary (3.2) of the next section.

3.2. High-Level Design

Based on the identification of design requirements (3.1) this section introduces a high-level design to CSP including its components. This design provides the foundation for two use case based design and prototype implementations. The first use case investigates CSP in closed domains (3.3) and the second in open domains (3.4).

3.2.1. High-Level Design Considerations

The proposed CSP approach is based on the design requirements discussed in section 3.1 of this thesis. It compromises of a third party service that interfaces with the different websites the user is browsing. This ensures that the independently hosted websites can offer CSP without a dependency related to the websites technology stack. From a user perspective a service approach to CSP provides a central access point ensuring profile scrutiny and control. Furthermore device and browser independence is ensured.

Following Figure 4 illustrates the notion of a third-party service based approach to CSP. The main elements are the Cross-Site Browsing Space consisting of websites that interface with a central third-party CSP service. The service offers a RESTful API that can interface with the websites to facilitate CSP based on the user's cross-site browsing within the Cross-Site browsing space. For simpler integration of CSP to the websites, website extensions are introduced to the approach.

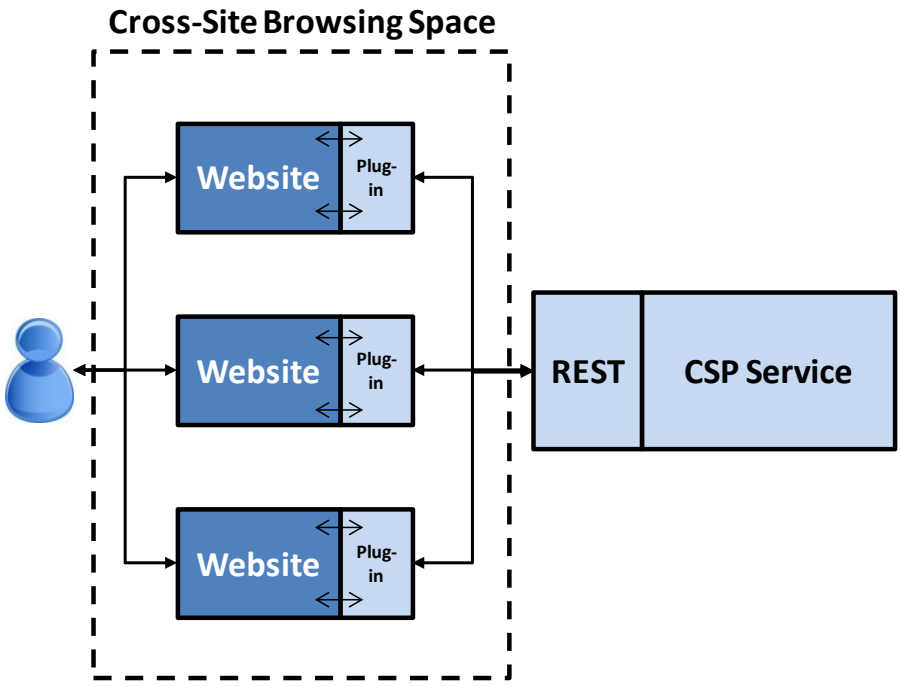


Figure 4 High-Level Service View

To ensure a third-party service approach is successful in providing CSP to different websites it is important to address interface/communication challenges. The interfacing mechanism between the different websites and the third-party service should facilitate constant interaction. This interaction should be generalised to avoid a limitation of the approach to a specific website design, content subject or technology stack. In summery the proposed CSP approach consists of three high-level architectural components (Figure 4):

1. A third-party service to facilitate accurate, user focused and privacy sensitive CSP.
2. A generic API (Application Programming Interface) independent from specific website designs, content subjects and technology stacks.
3. Website module extensions to allow simple and cost effective integration of CSP in a non-invasive manner.

The following section extends this initial design and provides details on the individual components within the architecture.

3.2.2. CSP Architecture Overview and Component Responsibilities

This section extends the introduced initial high-level design considerations in the previous section (3.2.1) towards a conceptual architecture. This architecture reflects the high-level design requirements (3.1) and provides the conceptual basis for two use case based design extensions and prototype implementations. The first use case focuses on CSP in closed domains (3.3) and a second use case focusing on CSP in open domains (3.4). The goal of the introduced architectural overview is to illustrate the responsibilities and capabilities of the different architectural components proposed.

All concepts introduced in relation to the design and architecture of CSP originates with the author of this theses. This includes the high-level design introduced in this section as well as the design, implementation and evaluation of the first (closed domain) use case. In relation to the second, more complex, use case (open domains) the author's sole contribution lies in the conceptual architecture, design and evaluation. The implementation of a prototype addressing the second (open domain) use case was supported by a Research Assistant. This involvement was mainly related to implementation aspects[67].

In the following the architecture (Figure 5) extending the initial high-level service view (Figure 4) is discussed. After this each component is introduced and described. This discussion includes the consideration of the high-level design requirements (Table 1 on page 64). The proposed architecture is used and extended throughout this thesis within two use cases. The specific use cases discussed in this thesis relate to open and closed domains.

[67] It involved the implementation of sign-in through OAuth, the utilisation of Zend Framework and the implementation of a WCMS extension for Drupal 7

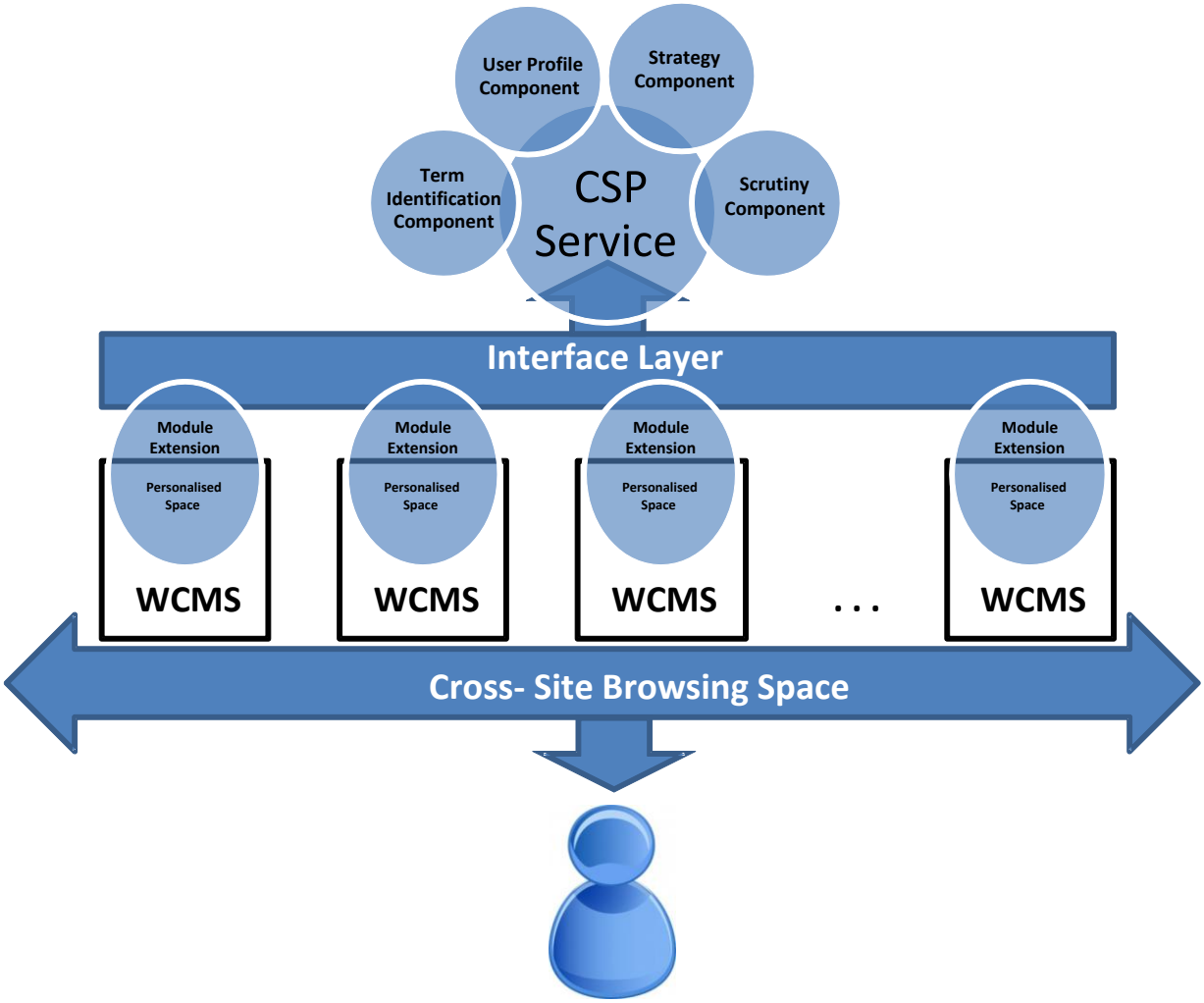


Figure 5 High-Level CSP Architecture

3.2.2.1. Term Identification Component

The term identification component of the CSP architecture has the responsibility to facilitate the identification of terms. The terms indicate the meaning of the underlying content from the websites within the cross-site browsing space. This identification can be a result of using already existing terms provided by the website or by using additional services/tools to identify terms. An additional responsibility of this component is to ensure the creation of a *shared conceptualisation* of the user's CSP enabled browsing space. This shared conceptualisation is represented as term space and is based on the content of the websites the user is browsing across. Once the terms are identified they can be used by the user profile component to represent the user's information needs. Furthermore, the term identification component requires means to facilitate continuous content updates to ensure

the shared conceptualisation of the user's cross-site browsing space is up-to-date. For this the term identification component needs access to the openly accessible content on the different websites. Within the interface layer of this architecture it is proposed that the website sends the content to the website that is to be used for CSP. This ensures the independence of the websites.

3.2.2.2. User Profile Component

The user profile component is responsible for the correct storage and aggregation of terms related to the content the user is browsing. These terms are provided by the term identification component and are stored together with the activity the user is conducting on the content. This activity can include mouse clicks, scrolls and keyboard presses. Together the terms and the activity form the base to create an informed decision by the strategy component. Furthermore the terms are used within the Scrutiny Component to enable the representation of the user's profile. The User Profile Component requires the identification of the user. This identification can either be provided directly by the website or by a sign-on mechanism, such as OAuth[68].

3.2.2.3. Strategy Component

The strategy component is responsible to provide an *informed decision*[69] to the website the user is currently browsing. An Informed decision assists a website to identify relevant links based on the users cross-site browsing. It implements a decision process that facilitates elements of the user's current user profile and relates this with the current website the user is browsing. The resulting *informed decision* allows a website to receive relevant information to provide adaptive guidance based on the user's overall information needs. The approach should implement techniques that ensure a fast and reliable response. For this the proposed approach should implement techniques that ensure fast and flexible calculation of an informed decision. Finally the strategy component has to ensure that the user's privacy is honoured. This is achieved by not passing any information related to the users cross-site needs to individual website. The design approach followed here is to allow the service to

[68] http://oauth.net/

[69] This thesis introduced the concept of an *informed decision* as a technical representation that is based on the user's cross-site browsing needs and that can be used by the website to provide Cross-Site Personalisation to the user.

provide information on the relevancy of links, but to not provide information on why the links are relevant. The informed decision is closely linked with website extension. These extensions allow the simple application of CSP based on the informed decision.

3.2.2.4. Scrutiny Component

The scrutiny component addresses requirements related to privacy, trust and control needs of the user. It allows users to access their user profile and to influence the profile. Furthermore the user can receive information on where and how the user related data was collected and used.

3.2.2.5. Interface Layer

The interface layer provides an abstraction from the specific implementation of the different websites within the user's cross-site browsing space. For this it implements a RESTful API (Advanced Programming Interface) to facilitate the communication between the CSP Service and the websites the user is browsing. A key responsibility of the API is to ensure that the interface with the CSP approach is device independent and does not depend on a specific technology stack. Furthermore, it should ensure fast and accurate interconnectivity between the interfacing websites and the CSP Service. The API can either interface with a website through bespoke integration or through provided website extensions.

The design requirements directly relating to the interface layer are simple interoperability and accurate and timely personalisations.

3.2.2.6. Web-based Content Management System Module Extensions

Web-based Content Management Systems (WCMS) module extensions allow a simple and non-invasive integration of CSP to existing website implementations. The responsibility of the WCMS module extensions is twofold: (1) To facilitate the communication between the website and the API of the CSP Service and (2) to provide non-intrusive personalisations to the user, within the website the user is currently browsing, based on an informed decision from the CSP Service. The introduction of CSP by providing non-intrusive assistance for the user and ensuring non-invasive integration in existing websites (during run-time) is an

important element of this research. Therefore it should be argued that WCMS module extensions are part of the high-level design of the introduced approach.

3.2.2.7. Cross-Site Browsing Space

The application of CSP to independently hosted websites introduces a Cross-Site Personalisation Browsing Space. Within the browsing space the user receives CSP through navigational assistance. The navigational assistance is enabled either through direct integration with the API of the CSP service or by using the above introduces website extensions. Based on the back-end integration of the CSP service with the website it is possible that the user decides when the CSP is enabled and when disabled. This allows the user to control the mechanism and ensures the user's privacy is honoured. The enabling/disabling of the service can either be facilitated though the user signing into the website or by directly signing into the service.

3.2.3. Summary

For overview purposes the introduced high-level requirements (3.1) are related to the above introduced high-level design component. Note following table:

Section	Concept Description	Component (Figure 5 page 68)	Identifier
3.1.1	A CSP approach should propose assistance that ensures the user is not hindered in their current browsing.	WCMS Module Extensions	[WUD-1]
3.1.1	A CSP approach should have a unified understanding of the user's browsing space by creating a shared conceptualisation.	Term Identification Component and User Profile Component	[WUD-2]
3.1.1	A CSP approach should apply a user modelling technique that does not interrupt the user's browsing. It should therefore be implicit by not requiring the user to explicitly participate in the identification, collection and management of information needs.	WCMS Module Extensions and Term Identification Component	[WUD-3]
3.1.1	A CSP approach should honour the user's	Strategy Component	[WUD-4]

	privacy needs by ensuring the different websites the user is browsing do not receive any information about the user cross-site information needs.		
3.1.2	A CSP approach should be able to utilise existing content models to create a shared conceptualisation across different websites.	Term Identification Component	[WCD-1]
3.1.2	A CSP should be able to use additional services and tools for term identification if necessary.	Term Identification Component	[WCD-2]
3.1.2	Depending on the domain (open or closed) the CSP approach should be able to identify terms of content that is added or updated to website as soon as possible.	WCMS Module extensions and Term Identification Module	[WCD-3]
3.1.3	A CSP approach should provide non-intrusive assistance techniques to ensure the user is not hindered in freely browsing.	WCMS Module Extensions	[WAD-1]
3.1.3	A CSP approach should affect the website as minimally as possible.	WCMS Module Extensions	[WAD-2]
3.1.3	A CSP approach should not negatively influence core areas of the website (loading time etc.).	WCMS Module Extensions	[WAD-3]
3.1.3	A CSP approach should be based on a design that allows simple integration and interfacing with existing websites.	WCMS Module Extensions	[WAD-4]
3.1.3	A CSP approach should ensure that the communication between the CSP approach and the independent websites is flexible and does not depend on the websites technology stack.	Interface Layer	[WAD-5]
3.1.3	A CSP approach should ensure device and browser independence.	Interface Layer	[WAD-6]

Table 2 High-level requirements related to design components

This section introduced a high-level design to CSP together with architecture components and their responsibilities. The purpose of this section was to provide a design abstraction that can be used within different CSP use cases. The following two sections introduce two distinct, but interrelated use cases. The first use case focuses on closed domains and the second use case focuses on open domains.

3.3. Cross-Site Personalisation in Closed Domains

The following section introduces a use case based design for applying CSP in closed domains. This proposed design is applied in the implementation (4.1) and evaluation (5.3) of a prototype applying CSP in a closed domain.

The structure of this section is as follows: First, the characteristics of a use case in a closed domain are discussed. After this the design considerations of a use case in a closed domain is introduced. This is followed by a discussion of the CSP architecture (based on Figure 5 in section 3.2.2). Finally, a summary is provided in which the specific elements of this use case are summarised in relation to the design requirements and the introduced approach in this use case.

3.3.1. Characteristics of a Closed Domains in the Context of CSP

It can be stated that the identification and subsequent assisting of users in cross-site information needs through CSP is challenging. The main reason for this is that the web is an open problem space in which a multitude of different websites addresses several different concerns of the users (Park and Kim, 2000). CSP use cases therefore have to be founded in actual information needs that the user would like to address and that require cross-site browsing.

To gain an initial understanding of the user's need for CSP an initial online user survey was conducted (5.2). This survey was conducted with computer researchers and experts and has led to the identification of a cross-site use case related to technical and/or research issues.

This initial investigation has led to the conclusion that there is a connection between the need the user wants to address and the type of website that can address this need. For example addressing challenges related to a technical implementation usually results in a cross-site browsing pattern, which involves structured corporate content (e.g. for manuals and technical description on enterprise hosted websites) and user-generated content (e.g. forums and Q&A websites). Based on this initial investigation CSP use cases within a closed domain can be categorised as **specific problem focused use cases** with the following characteristics:

a. A clearly defined problem, which usually contains questions of the type 'How to'.
b. Clearly defined terms to express the need of the user (partial user need expression).
c. Usually a pre-known set of websites to address the problem (closed domains based on mostly pre-known and trusted websites)
d. Predictable content types, such as structured technical manuals and user generated forum content.
e. Usually consist of individual unconnected, shallow and short cross-site browsing sessions.
f. Example real-world use case instance are among others: Customer Care, tasks related enterprise website federations (e.g. intranets).

The following section describes specific design requirements relate to the above introduced use case.

3.3.2. CSP Design Considerations for Use cases in Closed Domains

The introduced CSP approach relies on different high-level considerations (3.2.1) and high-level architectural components (3.2.2). This thesis addresses two specific instances of the overall architecture by introducing two distinct, but interconnected CSP use cases. This section discusses a use case design for CSP within a closed domain.

Design consideration for closed domains use cases are:

1. Build a *shared conceptualisation* across the different websites within the closed web browsing space.
2. Build/Use a content model based on a Taxonomy/Ontology of the semi-structured website content.
3. Use the Taxonomy/Ontology to train a statistical content model (e.g. by using a machine learning based text classifier).
4. Use the trained content model to calculate correlation between content of other websites with the pre-trained model.

5. Assign terms that have been identified by the trained model (correlating with the newly added content) with the other websites within the closed web browsing space.
6. Enable communication between the CSP Service and websites within the closed domains to facilitate an informed decision to assist the user in cross-site information needs.
7. Provide API endpoints that abstract the interaction to ensure device and web platform independence.
8. Ensure that all necessary information is exchanged between the CSP Service and the different websites.
9. Calculate an informed decision based on the user's profile and the current website interaction of the user.
10. Ensure the interaction between the website and the CSP Service is sufficient to enable CSP at the right moment.
11. Render the informed discussion on the webpage the user is currently viewing as non-intrusive visual guidance. This may include affecting different navigational structures within the websites, such as menu links, search result lists and forum entries.
12. Manipulate existing plug-ins/extensions to allow the rendering of an informed decision provided by the CSP Service.
13. Provide access to the user profile for the user.

The following section reflects the influences of the above-introduced design requirements on the different components of the high-level CSP architecture introduced in the previous section.

3.3.3. CSP Architecture Components and Responsibilities in Closed Domains

This section describes the different architectural components and responsibilities derived from the high-level design introduced in section 3.2.2. An important element of this use case is the requirement to generate a shared conceptualisation of the CSP enabled browsing space by using/generating a pre-extracted structure (e.g. taxonomy/ontology). This stage (the generation of a content model that can be used as a trained model for further term

identification in different web content) is technically not part of the architecture. In the following however the process of acquiring and using a trained model described as part of the term identification component.

3.3.3.1. Term identification Component

The identification of terms across different content sets is a challenging endeavour with several different approaches discussed in the State of the Art of this thesis. Based on the specific use case discussed here, this design instance requires a pre-extracted representation of a website to start the process. This representation can be used to extract meaning from the content of other websites. To ensure a high correlation between the trained model and newly added content the subject covered by content should be similar across all websites within the CSP enabled browsing space. As noted in the introduction this pre-phase is a design time assessment, meaning that the creation of a model and the assigning of the extracted terms to the different webpages is static. Any extensions, updates or modifications to the content model may trigger a re-assessment of all content across all websites within the CSP enabled browsing space. Tools that can be used to train a content model are typically machine learning tools (Witten and Frank, 2005; Witten et al., 1999). This class of tools relies on a given corpora that includes assigned pre-annotated keywords. The correlation between the content and the keywords allows the creation of a trained model that enables the statistical calculation of how strong the correlation of newly added content to the existing model is. Based on the closed domain nature of this use case, the identified terms can be assigned to the web pages directly through either meta-tags or other means of assigning keywords to content.

In this use case an initial taxonomy should be created to allow the training of a statistical model. Research by Sah and Wade (2010) provides an approach to extract a term model. For this approach to work the initial website should be semi-structured as XML representation. A specific representation discussed in the State of the Art is DocBook. It is usually applied in corporate enterprise websites that underlie a strict editorial structure. For this use case the approach can be used to extract a representation from an initial semi-structured website. This representation can assist the training of a model allowing the annotation of websites

that do not underlie a structure (e.g. based on user generated content such as blogs and forums).

3.3.3.2. User Profile Component

The user profile component is responsible to store terms related to content within the CSP enabled browsing space. In addition to the storage of terms additional evidence can be attached to the keyword such as a counter, time (related to the reading time of related content) and other activities (such as clicks, scrolling etc.). Based on the pre-extracted and assigned keywords the shared conceptualisation of the closed browsing space is finite and therefore it is possible to pre-populate the user's profile with the full set of keywords by the Term identification Component.

3.3.3.3. The Strategy Component

The strategy component is responsible for the calculation of an informed discussion that is based on the user's profile (as a representation of the user's current information needs) and the current web pages the user is currently browsing. In this closed domain use case the strategy component is provided with a finite set of terms that represent the information on the different web pages within the closed browsing space. The main function the strategy component is to correlate the existing representation of the user's information needs (user profile) with the possible navigational decisions the user can make on the current webpage. The result of this correlation is an informed decision, provided by the strategy component, and which allows the website to influence the user's navigational decision through non-intrusive visual guidance. Based on the closed nature of the domain the terms representing the user's cross-site browsing interest can be exposed to the websites. In that case privacy sensitivity is provided by not exposing the source of the terms.

3.3.3.4. The Scrutiny Component

This component allows the user to access a visual representation of the information the CSP Service has stored about the user. It therefore can be seen as a visual extension of the user profile. In closed domains the user profile visualisation consists of all terms the CSP Service has assigned to the user's profile and that relate to the user's current information needs.

3.3.3.5. The Interface Layer

The interface layer enabled the communication between the CSP Service and the different websites the user browses. The communication is based on distinct API endpoints based on Representational State Transfer (REST)[70] to ensure a high-level of abstraction.

3.3.3.6. WCMS and personalised spaces

The personalised space introduced by the CSP approach is influenced by an informed decision provided by the CSP Service. Due to the use of WCMS module extensions the influence of the CSP approach is limited to the elements/functions that the applied WCMS extension can influence. Due to the non-intrusive notion of CSP, introduced with this research, the influence by the CSP Service is limited to visual clues that assist the user in making a decision on what navigation step to take next. The personalised space therefore does not limit the decision space of the user, but adds a visual layer. Based on the RESTful API it is possible to use an informed decision, provided by CSP Service, for more intrusive scenarios. The investigation of more intrusive influences within the user's personal browsing space is not part of this research.

3.3.3.7. Closed Domain Browsing Space

The closed web browsing space of this use case is based on the notion that all web pages are known at design-time. This closed domain can be extended flexibly by adding websites. However, the added websites need to be related to subject matter of the website within the closed web browsing space. This restriction is due to the trained model, which is used to analyse and annotate all web pages within the closed web browsing space.

3.3.4. Summary

The section above introduces an approach to apply CSP in a closed web browsing space. The distinct characterises of this browsing space were described after which the design requirements were identified.

[70] http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm

The description of closed domains was linked to the categorisation of a use case within such domains and introduced as specific problem focused use cases. The characterises of a closed domain use case introduced are a clearly defined problem, clearly defined terms to express the need of the user (partial user need expression), usually a pre-known set of websites to address the problem, predictable content types (such as structured technical manuals and user generated forum content). Furthermore such use cases usually consist of individual unconnected, shallow and short cross-site browsing sessions. Use case instances identified are among others Customer Care, purchase and tasks related enterprise website federations (e.g. intranets).

The study of CSP in closed domains was chosen to enable a more controlled study leading towards a prototype installation (4.1) and evaluation (5.3) within a closed domains use case. Based on the iterative nature of this research, the outcomes of this initial closed web study inform the design, development and evaluation of a CSP approach within open, and therefore more complex, domains.

The following table is based on the initial requirements table introduced in the summary of section 3.2 (Table 1 on page 64). The table is extended with a column indicating what approach has been proposed by the design of CSP in closed domains.

Section	**Concept Description**	**Component** (Figure 5 page)	**Proposed Design Approach**	**Identifier**
3.1.1	A CSP approach should propose assistance that ensures the user is not hindered in their current browsing.	WCMS Module Extensions	Non-Intrusive visual guidance through link annotation.	[WUD-1]
3.1.1	A CSP approach should have a unified understanding of the user's browsing space by creating a shared conceptualisation.	Term Identification Component and User Profile Component	The user profile consists of pre-extracted terms related to the content of links the user clicks.	[WUD-2]
3.1.1	A CSP approach should apply a user modelling technique that	WCMS Module Extensions and	The implicit detection of	[WUD-3]

	does not interrupt the user's browsing. It should therefore be implicit by not requiring the user to explicitly participate in the identification, collection and management of information needs.	Term Identification Component	information needs is based on the content the user accesses.	
3.1.1	A CSP approach should honour the user's privacy needs by ensuring the different websites the user is browsing do not receive any information about the user cross-site information needs.	Strategy Component	The CSP service provides an *informed decision* consisting of a list of terms. Each term represents a cross-site browsing need. The website receives no information about where the terms originated.	[WUD-4]
3.1.2	A CSP approach should be able to utilise existing content models to create a shared conceptualisation across different websites.	Term Identification Component	The shared conceptualisation is based on the initial structural analysis of the content to apply a machine learning text classifier based approach.	[WCD-1]
3.1.2	A CSP should be able to use additional services and tools for term identification if necessary.	Term Identification Component	Proposed as extension to the second prototype.	[WCD-2]
3.1.2	Depending on the domain (open	WCMS Module	Based on a text	[WCD-3]

	or closed) the CSP approach should be able to identify terms of content that is added or updated to website as soon as possible.	extensions and Term Identification Module	classifier approach introduced in this use case the classification model can be used to identify the correlation of the new/updated content with the existing term space.	
3.1.3	A CSP approach should provide non-intrusive assistance techniques to ensure the user is not hindered in freely browsing.	WCMS Module Extensions	The assistance is proposed as link annotation to assist the user navigation. The non-intrusive nature is based on applying link annotation to cross-site relevant links.	[WAD-1]
3.1.3	A CSP approach should affect the website as minimally as possible.	WCMS Module Extensions	By using WCMS customisable module extensions a website can be influenced without significantly affecting it.	[WAD-2]
3.1.3	A CSP approach should not	WCMS Module	The website is	[WAD-3]

	negatively influence core areas of the website (loading time etc.).	Extensions	responsible for the rendering of the informed decision. The decision is made within the CSP Service.	
3.1.3	A CSP approach should be based on a design that allows simple integration and interfacing with existing websites.	WCMS Module Extensions	Facilitated through WCMS module extensions and an API based interface layer.	[WAD-4]
3.1.3	A CSP approach should ensure that the communication between the CSP approach and the independent websites is flexible and does not depend on the websites technology stack.	Interface Layer	Facilitated through an (RESTful) API based interface layer	[WAD-5]
3.1.3	A CSP approach should ensure device and browser independence.	Interface Layer	Facilitated through an (RESTful) API based interface layer	[WAD-6]

Table 3 Closed Web Design Requirements

In the following section the application of CSP in an open domain is discussed. This is then followed by an overall summary of this design chapter.

3.4. Cross-Site Personalisation in Open Domains

The following section introduces a use case based design for applying CSP in open domains. This proposed design is applied in the implementation (4.2) and evaluation (5.5) of a prototype applying CSP in an open domain.

The structure of this section is as follows: First, the characteristics of a use case in an open domain are discussed. After this the design considerations of a use case in an open domain introduced. This is followed by the discussion of the CSP architecture (based on Figure 5 in section 3.2.2). Finally, a summary is provided in which the specific elements of this use case are summarised in relation to the design requirements and the introduced approach in this use case.

3.4.1. Characteristics of a Closed Domains in the Context of CSP

Analogue to the investigation and introduction of closed domain use cases in the previous section (3.3), this section introduces considerations related to use cases in open domains. These cases can be considered as higher in complexity due to the wide and open web space considered. This complexity is based on a more flexible CSP browsing space in terms of website members and content updates. In a closed domain the amount of websites within the CSP browsing space are limited and the content updates usually controlled.

Based on this initial investigation open domain use cases for CSP can be categorised as **general problem focused use cases** with the following characteristics:

a. An unclear problem space in which the user may gather information as part of a higher and not yet fully formed cross-site information needs.
b. User is not yet fully aware of what terms to use to identify the need.
c. Usually unknown set of websites to address the need. (Open domains based on mostly unknown websites)
d. Unpredictable content types.
e. Usually consists of many different interconnected cross-site browsing sessions within different time frames.

f. Example real-world use cases are among others: Research and planning, news, hobbies and activities etc.

The following section describes specific design requirements relate to the above discussed use case considerations.

3.4.2. Design Considerations for CSP Use cases in Open Domains

This section relates to use cases introduced as open domains. The design requirements introduced in the previous section (3.3), related to closed domain use cases, and are used as foundation for the following design requirements. Therefore the following design requirements can be seen as modification/iteration of the previously introduced requirements for closed domains (3.3.2).

The following design consideration for open domain use cases are introduced as follows:

1. Identify meaning of website content by using an open and external API based service to allow the identification of content terms during run-time.
2. Ensure the identification of content meaning is applied to newly added or edited content.
3. Use terms extracted from external service to populate user profile to ensure flexible extraction across different subject domains.
4. Index content to ensure term and content are matched in the CSP Service.
5. Assign users browsing behaviour (e.g. clicks, scrolls etc.) to the user profile to identify the relevancy of the extracted terms.
6. Use content index and the user profile to generate an informed decision in run-time.
7. Ensure communication between websites and CSP approach is sufficient for real time informed decisions.
8. Provide WCMS extension module that enables non-intrusive integration to WCMS based websites.
9. Ensure WCMS extension provides non-intrusive personalised visual assistance based on the informed decision from the CSP Service.

The following section reflects the influences of the above introduced design requirements to the different components of the CSP architecture.

3.4.3. CSP Architecture Components and Capabilities

This section describes the different architectural components and responsibilities in open domain use cases. This section is influenced by the architecture components and responsibilities introduced in the previous section in relation to closed domain problems. Based on the open web nature of this use case applying CSP in an open domains requires a more open and flexible approach in the application of understanding content. Furthermore run-time considerations have to be addressed. The understanding of content should be simple and fast due to websites often updating and adding content.

3.4.3.1. Term identification Component

The term identification component has two responsibilities in an open domain use case. The first responsibility is related to identifying the meaning of content extracted from website through external services. And the second relates to storing a representation of the web content together with the identified terms for further usage by other components. In an open domain use case the restrictions of a trained model are lifted by allowing the term identification component to manage one or many external services to provide the extraction of meaning from web content. Once a new website is added to the CSP browsing space (by deploying the WCMS extension module or directly interfacing with the API) the term identification service starts to create an index (or stored representation) of the publically available content sent by the newly added website. This representation is extended based on terms that are extracted from external services and which identify the meaning of the underlying content.

3.4.3.2. User Profile Component

The user profile component for open domain use cases is similar to closed domain use cases. Two main extensions are the usage of additional user behaviours and the storage of where the behaviour has originated. Both responsibilities serve two different purposes. The first is to support the strategy component to generate an informed decision. The second is to allow scrutiny for the user and to understand where user data has originated.

3.4.3.3. Strategy Component

The strategy component, which is responsible for the calculation of an informed discussion, should implement a fast, scalable and accurate technique to generate an informed decision based on the users cross-site information need. In an open domain the term space is potentially unlimited (not finite like in a closed domain use case). Therefore the introduced approach to generate an informed decision should not be based on term matching or mapping. However, it should include the terms as a valid representation of the user's cross-site information needs. For this, techniques related to Information Retrieval and Recommender Systems are most appropriate due to their wide and successful usage in different applications on the open web. In relation to privacy sensitivity the strategy component ensures that the informed decision is based providing link relevancies to the requesting website. The links are in-bound therefore only within the requesting website. Furthermore no terms or other sources of information are added to the informed decision provided by the CSP service.

3.4.3.4. Scrutiny Component

The scrutiny component in an open domain use case has the same responsibilities as in a closed domain use case. The main difference is that the user may request additional information on how the information was used in the open web. This is a result of the user possibly applying different trust levels in an open domain use case compared to a closed domain use case. The user might have a lower level of trust due to not knowing all the websites within the open web CSP browsing space. Furthermore the user might want to control what information the different websites are allowed to use. Both techniques, providing user data usage information and user data control, are not part of this thesis, but are discussed in the future work section.

3.4.3.5. Interface Layer

The interface layer, enabling communication between the third-party service and the different websites the user browses, has similar responsibilities as in the closed domain use case. Due to the necessity of the open web nature requiring integration with different

website technologies it is important to ensure the level of abstraction is maintained to avoid the RESTful API being limited to a specific subject, technology or device.

3.4.3.6. WCMS and personalised spaces

The Web-based Content Management Systems (WCMS), as a component of the overall CSP approach is based on a continuous development approach. In the context of this thesis WCMS extensions should ensure useful integration of CSP at run-time (this includes backend and content related matters, such as sending content for indexing to the service). Furthermore, interfacing with the CSP Services API allows third party developers to extend and add WCMS module extensions.

3.4.3.7. Open Domain Browsing Space

The open web browsing space is based on the notion that all webpages are not known at design-time and that the browsing space constantly changes based on the amount of websites adopting the CSP approach. Furthermore the content base may change rapidly and therefore the CSP Service should ensure that any informed decision is accurate based on up-to-date non-intrusive navigational guidance.

3.4.4. Summary

The section above introduces an approach to apply CSP in an open web browsing use case. The distinct characterises of this browsing space were described after that the design requirements were introduced. The requirements subsequently informed the introduction of an open domain use case CSP architecture and its components.

The description of open domains was linked to the categorisation of a use case within that domain and introduced as general problem focused use cases. The characteristics of an open domain use case introduced are an unclear problem space in which the user may gather information as part of a higher and not yet fully formed cross-site information needs, the user is not fully aware of what terms to use to identify the need, usually unknown set of websites to address the need, unpredictable content types and usually consists of many different interconnected cross-site browsing sessions with different time frames. Use case instances identified are among others research and planning, news, hobbies and activities.

The study of CSP in open domains was based on the outcome of the first use case. The main difference between both use cases is extensions to the WCMS module extensions and a more flexible (IR-based) approach to generate an informed decision. Furthermore the term identification component within this second use case based design has been extended to interface with a third-party term identification service. This enables a higher level of flexibility in terms of webpage content usage. It should be noted that the second use case based design can be equally applied within closed domains. It furthermore can be argued though that the usage of a specific content description model, as introduced in the closed domains, may be necessary to maintain data security (within closed enterprise website federations) by not exposing content to a third-party term identification services.

In relation to the overall design requirements following requirements were addressed by this use case design:

Section	**Concept Description**	**Component** (Figure 5 page)	**Proposed Design Approach**	**Identifier**
3.1.1	A CSP approach should propose assistance that ensures the user is not hindered in their current browsing.	WCMS Module Extensions	Non-Intrusive visual guidance through link annotation. The annotations include a link relevancy indication.	[WUD-1]
3.1.1	A CSP approach should have a unified understanding of the user's browsing space by creating a shared conceptualisation.	Term Identification Component and User Profile Component	The user profile in an open domain use case consists of terms that are identified by the term identification component through facilitation of	[WUD-2]

			additional external services. The terms space is therefore open and not limited to a trained model.	
3.1.1	A CSP approach should apply a user modelling technique that does not interrupt the user's browsing. It should therefore be implicit by not requiring the user to explicitly participate in the identification, collection and management of information needs.	WCMS Module Extensions and Term Identification Component	The implicit detection of information needs is based on the browsing behaviour of the user. The content related terms are within the CSP Service and are not shared with the websites.	[WUD-3]
3.1.1	A CSP approach should honour the user's privacy needs by ensuring the different websites the user is browsing do not receive any information about the user cross-site information needs.	Strategy Component	The informed decision is represented as links annotations. Each link including calculated cross-site relevancy. No terms are transferred. The website receives no information about how the informed decision was created.	[WUD-4]
3.1.2	A CSP approach should be able	Term Identification	Shared	[WCD-1]

	to utilise existing content models to create a shared conceptualisation across different websites.	Component	Conceptualisation is based on the using external term identification services. The content of CSP enabled websites is stored within the CSP Service and indexed based on the identified terms.	
3.1.2	A CSP should be able to use additional services and tools for term identification if necessary.	Term Identification Component	Not applicable in open domain use case.	[WCD-2]
3.1.2	Depending on the domain (open or closed) the CSP approach should be able to identify terms of content that is added or updated to website as soon as possible.	WCMS Module extensions and Term Identification Module	Term identification is based on an external term identification services. The term identification component is designed modular to ensure the interchanging and interconnecting of term identification services if necessary.	[WCD-3]
3.1.3	A CSP approach should provide	WCMS Module	The	[WAD-1]

	non-intrusive assistance techniques to ensure the user is not hindered in freely browsing.	Extensions	personalisations are visualised by the website through WCMS module extensions. To ensure consistence across the different websites the module extensions is confined to a specific area of the website (e.g. a link menu structure).	
3.1.3	A CSP approach should affect the website as minimally as possible.	WCMS Module Extensions	WCMS module extension can be deployed without re-designing the website.	[WAD-2]
3.1.3	A CSP approach should not negatively influence core areas of the website (loading time etc.).	WCMS Module Extensions	The information exchanged between the website and the service is kept as minimal as possible (user behaviour information and link lists) to ensure the websites	[WAD-3]

			performance is not influenced.	
3.1.3	A CSP approach should be based on a design that allows simple integration and interfacing with existing websites.	WCMS Module Extensions	The module extensions can be deployed to the websites and enabled during runtime.	[WAD-4]
3.1.3	A CSP approach should ensure that the communication between the CSP approach and the independent websites is flexible and does not depend on the websites technology stack.	Interface Layer	The interface layer is based on a RESTful API to enable communication with different websites.	[WAD-5]
3.1.3	A CSP approach should ensure device and browser independence.	Interface Layer	The generic API provides independent from device and website specific technologies by only concentrating on user specific information	[WAD-6]

Table 4 Open Web Design Requirements

3.5. Conclusions

This chapter introduced a high-level design to CSP (3.2). This high-level design was influenced by the introduction of design requirements (3.1) derived from the State of the Art chapter of this thesis. This design introduced several different design considerations (3.2.1), a high-level architecture (Figure 5 on page 68) and relevant architectural components (3.2.2) that a CSP approach require to assist users in addressing cross-site information needs. This high-level design was then subsequently applied to inform the design of two specific CSP use cases. The first use case related to closed domains (3.3) and the second use case related to open domains (3.4). Both use cases were described including the introduction of use case characteristics, design requirements and architecture component. Even though both use cases were introduced as separate use cases it should be noted that they are interconnected. The main differences in design, such as the term identification through a third party service and the IR-based approach to informed decision creation (strategy component) is a result of the different problem spaces each use case addresses.

Within a closed domain the problem space is introduced as more specific and content updates are more structured and not as rapid. Therefore it was necessary to extend the design for open domain based use cases by allowing a more flexible term identification process and a more rapid and flexible approach to create an informed decision.

Furthermore, considerations in relation to changes in the size of the CSP browsing space were addressed. It can be argued that the adding of websites within a closed domain does not happen often and if it does it is mostly planned and structured. Within an open domain the adding of websites to the CSP browsing space should be handled more flexible. To achieve this, the initial design and responsibility of WCMS module extensions for the closed domain were extended.

In relation to the high-level design requirements introduced at the beginning of this chapter (3.1), it can be concluded that all requirements are addressed by the two introduced use case based designs (Table 4 on page 93). It furthermore can be argued that both

interconnected use case designs address the challenges and gaps identified in the State of the Art based review.

In the following two chapters first the implementation (chapter four) of both designs is introduced followed by the evaluation (chapter five).

4. Implementation

The previous chapter described the design of a CSP approach that can assist users addressing in information needs that span independently hosted websites. A high-level design was introduced together with tow specific design instances related to two distinct, but interconnected use cases. The first CSP use case illustrates CSP in closed domains, the second in open domains. For both use cases a design was introduced in connection with the discussion of design requirements derived from the State of the Art review.

This chapter describes the implementation of two prototypes. Each prototype addresses one use case. The development process was iterative with the outcome of the first prototype implementation (closed domain) informing the implementation of the second prototype implementation (open domain).

This chapter is organised as follows. In the following section 4.1 the implementation of a CSP prototype in closed domains is described. After this in section 0 the implementation of a second CSP prototype is discussed. This second prototype introduces CSP in open domains and is built based on the evaluation outcomes of the first prototype. Both sections are structured equally.

4.1. Prototype Implementation for Cross-Site Personalisation in Closed Domains

This section illustrates the implementation of a prototype to assist users in cross-site information needs within a closed domain. This section is organised as followed: First the characteristics of closed domains are described (4.1.2). This description is based on the design chapter (3.3.1). This is followed by introducing a use case example together with use case requirements (4.1.2). The example use case is consistent with the evaluation of this prototype implementation discussed in the evaluation chapter (5.3). After this, content considerations are introduced (4.2.3). The technological architecture and its components are introduced (4.1.4). Finally, a process overview (4.1.5) and a summary (4.1.6) are provided.

4.1.1. Closed Domain Characteristics

Based on the initial use case discussion in the design chapter (3.3.1 page 74) of this thesis following main characteristics of a closed domain summarised as following:

- Only users that are authenticated are able to access the content (i.e. the user knows and trusts the websites due to being registered).
- The information needs of the user are usually focused e.g. fact/information finding and related to a specific problem.
- Closed domains are often found in large-scale enterprise level websites and website federations.
- Changes to the websites structure or content are usually applied planed (based on editorial and corporate design limitations).

In the following use case example the application of CSP in a closed domain is discussed together with content and implementation details specific to the introduced use case.

4.1.2. Use case Example and Pre-Requisites

This section discusses an example use case within a closed domain. The specific area chosen for this use case is Online Customer Experience (OCE)[71] (Kobsa et al., 2001). This area was chosen based on feedback received in an initial user survey (5.2).

In this example use case, a user browses to different websites related to technical product information. The first website hosts structured enterprise content based on Symantec's Norton 360 product range[72]. The second website hosts unstructured forum content also related to the Norton 360 product range. Figure 6 illustrates CSP within this use case.

[71] The selection of this use case is based on an initial user survey indicating usefulness of CSP in customer care use cases.

[72] The content used for this use case was kindly provided by Symantec and is on live Symantec websites.

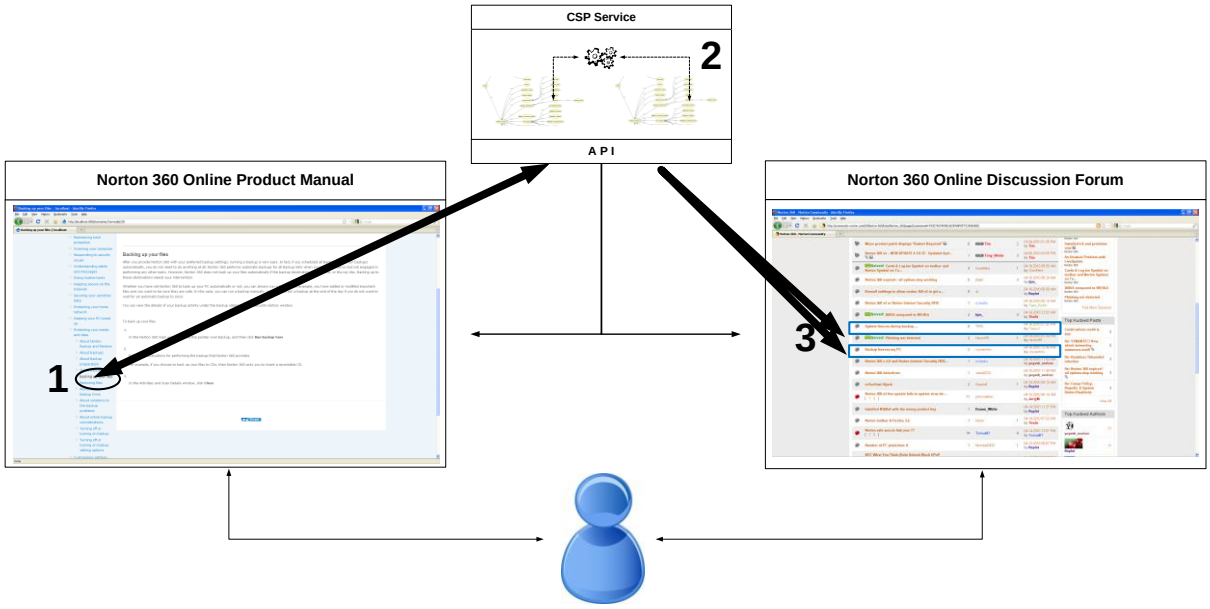


Figure 6 High-level Example Use case of CSP in a Closed Domain Environment[73]

The cross-site browsing process illustrated in Figure 6 is as follows:

1. The user browses structured (e.g. corporate) content, such as a product manual (in this example the Norton 360 product manual). The CSP Service receives activity information from the website to provide an informed decision based on the user's indicated information needs.
2. The CSP Service adds pre-extracted content information (e.g. terms) to the existing user's profile based on the user's interactions. The user profile is used to generate an informed decision to assist the user on a second website.
3. The second website receives an informed decision and uses this to render visual assistance for the user and based on the user's cross-site browsing need.

Applying CSP to the above-illustrated use case in closed domain underlies following requirements:

1. An initial representation of the content subject within the use case should be presented (e.g. specific product information or learning subject) in a structured representation (e.g. DocBook, Taxonomy or Ontology).

[73] Even though the arrows indicate a one way direction, the CSP approach is implemented to support users browsing in both directions.

2. The provided content representation should include metadata representing the meaning of the content. If no meta data is provided it should be extracted either through simple methods, such as TF-ID (Kumaran and Allan, 2008) or through more sophisticated approaches, such as introduced by Sah and Wade (2010) by utilising ontology creation and reasoning.
3. The content set, together with the corresponding terms, may need to be manually validated for double entries and misleading metadata.
4. A text classifier tool can be used to create a trained model of the content together with the related terms. This model is subsequently used to identify terms in newly added content/websites.
    a. The initial content set together with the terms needs to be transformed into a format readable by a text classifier (e.g. ARFF in the case of Weka[74]).
    b. To ensure the model resulting from text classification is accurate newly added content/websites should be related to a similar subject matter (Riloff and Lehnert, 1994).
5. The newly identified terms have to be added to the content of the websites.
6. To facilitate CSP the existing websites can interface with the CSP Service API through custom-built WCMS extension modules or by directly communicating with the API.
7. The websites within the closed CSP enabled domain can enable user identification e.g. through sign-in. This enables access to the content and allows the identification of the user for CSP.
8. CMS module extensions can be used to interface with the CSP Service and to enable personalised visual assistance of the informed decision provided by the CSP Service.

The following section discusses specific implementation details. First content considerations (4.1.3) are discussed after which the technological architecture is presented (4.1.4).

4.1.3. Content Considerations

The content for this use case based prototype implementation consisted of two content sets. For evaluation purposes both content sets had to be prepared for CSP across two

[74] http://weka.wikispaces.com/ARFF+%28stable+version%29

websites. This preparation included the creation of a shared conceptualisation of the content based on term extraction and the importing of content to the two different websites implemented as Drupal based WCMS. The following section describes both the preparation and the importing of the content used for this prototype implementation.

The first content set consisted of highly structured product manuals related to the Norton 360 product range, the second related to forum based user generated content. Both content sets had to be prepared for usage within this use case. The overall goal of this preparation was twofold: (1) To train a text classifier (based on the semi-structured content set) to identify correlations with the second unstructured content set and (2) to import both content sets to individually hosted websites for evaluation of the CSP approach. This preparation had to be performed at design time due to the impact of importing content to a website. For this prototype implementation two new websites were set-up. This however is not necessary if the website is already set-up. In this case it is only necessary to add the identified terms to the content either by adding meta-tag elements or by using WCMS specific modules responsible for metadata and taxonomy generation.

To train a text classifier based on an existing content set a structured representation had to be extracted from the DocBook representation of the semi-structured content that was provided by Symantec. DocBook provides a standard and structured representation as XML[75] structure. The DocBook representation allowed the extraction of a concept ontology based on an automatic transformation process introduced by Sah and Wade (2010). The resulting ontology allows the representation of DocBook as OWL based ontology with each section as individual (Figure 7).

[75] http://www.w3.org/TR/REC-xml/

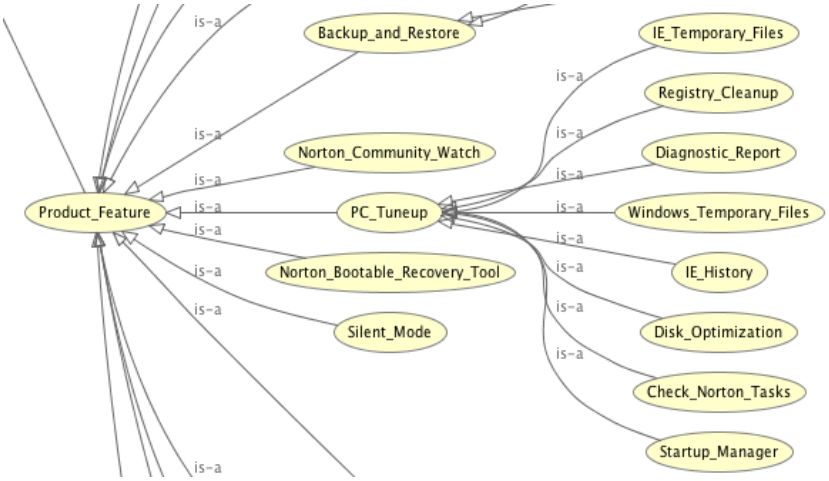


Figure 7 Concept Ontology (excerpt) based on DocBook extraction

The main goal of extracting a structured representation was to prepare the content for its usage in a text classifier. Each individual in the ontology represents a concept. For each concept, data properties were set and used for term extraction. The data property resembles the name of the individual. Furthermore each individual holds a unique reference id to the related content snippet. For text classification the data mining tool Weka[76] was used. To fulfil the requirements of Weka (the data needs to be stored in the file format ARFF or XRFF)[77] a transformation script was written in Java by using the Jena ontology-parsing library (Jena Semantic Framework v.2.63)[78]. Following process overview summarises the approach taken:

1. The initial enterprise content was converted from DocBook into an Ontology representation.
2. This representation included identified terms based on the concept description of the ontology representation.
3. A text classifier readable file was created (in this case a weka compatible ARFF file) with the content and the associated concept terms (e.g. backup, safe recovery etc.).
4. A text classifier was used to create a model representation.
5. This representation was used to classify additional content that had no metadata assigned.

[76] http://www.cs.waikato.ac.nz/ml/weka/

[77] http://weka.wikispaces.com/Text+categorization+with+WEKA

[78] http://jena.sourceforge.net/

6. The content was imported into a WCMS by using ontology parsing libraries and scripts to ensure the content structure was maintained.

The second content set was based on user-generated forum content. This content was retrieved by crawling the online Symantec Norton 360 User Forum[79] with the Heratrix[80] web crawler. Overall approx. 10.000 forum entries were indexed. Most title fields had fuzzy descriptions therefore the use of title fields or other description could not be used to identify concept terms. The resulting ARC[81] file was transformed through scripts to prepare for classification. The trained classifier allowed the assigning of concept terms to the newly extracted content resulting in a shared conceptualisation (in the form of a term space) across both closed corpus content sets.

In a second step both content types were imported into websites implementing a Drupal WCMS. The structured manual based content was imported to enable structured navigation based on the initial structure provided in the DocBook representation. To import the content different PHP based scripts were developed. These scripts allowed automatic import including metadata terms. Furthermore, for the second website representing the user generated content within forums inbuilt content import[82] and taxonomy modules[83] were used.

The following section describes the technological architecture of the CSP approach.

4.1.4. Technological Architecture and Components

This section describes the technological architecture required for CSP in a closed domain. The architecture is based on the high-level architecture introduced in the design chapter of this thesis (Figure 5). In the following the technical architecture (Figure 8) and its components is presented.

[79] http://community.norton.com/t5/Norton-360/bd-p/Norton_360

[80] http://crawler.archive.org/

[81] http://en.wikipedia.org/wiki/ARC_%28file_format%29

[82] http://drupal.org/project/node_import

[83] http://drupal.org/project/module_taxonomy

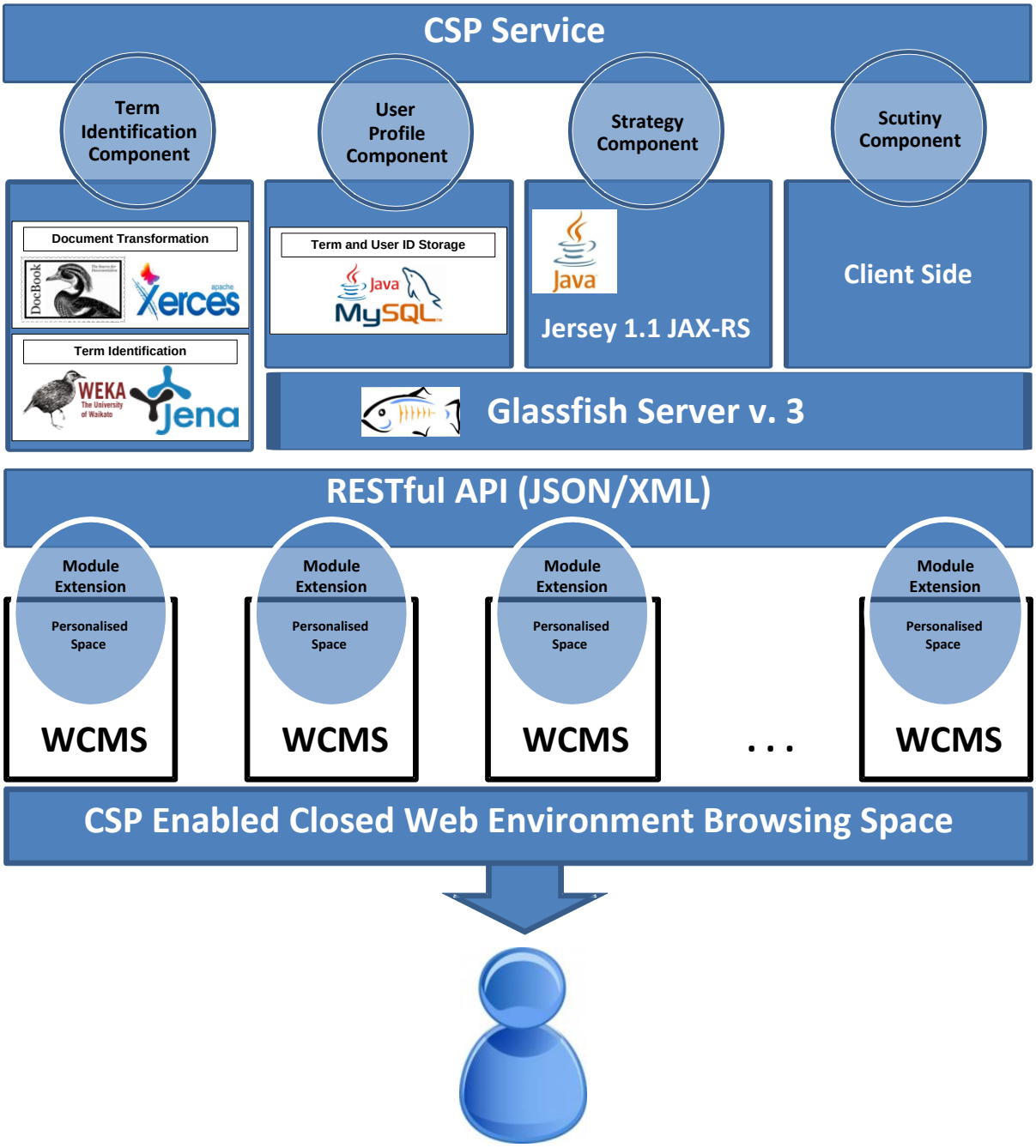


Figure 8 Technical CSP Architecture for Closed Domain[84]

Following sections discuss the implementation of the individual components necessary to address this use case with the introduced CSP approach.

[84] The implementation was conducted in JAVA.

4.1.4.1. Term Identification Component

The term identification is conducted in design-time including the identification of initial terms, the creation of a content model with a text classifier and the usage of the trained classifier to identify correlations between the already identified terms based on semi-structured content and using this model for the identification of terms in additional (unstructured) content. The process applied to identify terms is described in the content preparation section above.

4.1.4.2. User Profile Component

The user profile component is implemented as MySQL database[85]. For each user a unique user ID is generated that matches the user ID of the Drupal deployment. The user ID is used to send and receive the user specific terms stored in the user profile. The user ID is identified through the CSP module extension using the Drupal *hook_user*[86] API.

4.1.4.3. Strategy Component

In closed domains the strategy component enables the sending of the terms within the user's profile. This allows the matching of the pre-extracted terms with the terms related to the webpage links. The strategy component passes the terms stored in the user profile component to the website. For performance reasons the websites client conducts the matching of the links prior to page rendering.

4.1.4.4. Scrutiny Component

In this first prototype websites can request the term model of the user to display it to the user.

4.1.4.5. WCMS Client Integration

This section describes both the integration of content based on a shared conceptualisation (introduced in section 4.1.3 of this chapter) into a Drupal WCMS implementation and the extension of the WCMS based websites to facilitate CSP. For the use case prototype

[85] http://www.mysql.com/

[86] http://api.drupal.org/api/drupal/developer%21hooks%21core.php/function/hook_user/6

implementation Drupal (version 6.14[87]) was used. This decision is based on a State of the Art survey conducted prior to this integration (appendix on page 226). At the time of the investigation Drupal was identified as being widely distributed and advanced in providing a flexible and extendible module framework.

The importing of content was focused on two different content sets into two separately hosted websites (both implementing Drupal 6.14). The first website implementing a structured, menu driven website (based on the individual chapters and sections of the DocBook based content structure). The second website implemented a forum based website. The script was required in the first website due to building the menu structure in a recursive manner. The main content was imported in both websites with the help of the *node_import module*[88] and the *taxonomy module*[89].

Following steps were required to import content to Drupal based WCMS and to enable CSP. The example content sets are structured and unstructured (user generated) and CSP is enabled across each content set hosted on separate websites.

1. Save the semi-structured content in a representation that allows importing to a website and that maintains the content structure (e.g. XML).
2. Save the unstructured content representation in a file that relates the content with the classified terms and that allows importing to a website (e.g. CSV file).
3. Use provided modules within Drupal to import content. In the case a structure should be replicated the usage/generation of a custom script may be required.
4. Import content on both websites that may include manual adjustments.
5. Ensure the terms are related to the content via a taxonomy that allows the WCMS based website to related content with the same terms.
6. Index the website to allow searching. For more effective search ensure the taxonomy terms are assigned to the search index.

[87] At the time of implementation Drupal 7 was available although most module extensions were not yet compatible.

[88] http://drupal.org/project/node_import

[89] http://drupal.org/project/module_taxonomy

7. Deploy and enable custom built CSP module extensions for Drupal to facilitate third-party communication and hooking to necessary function calls within the WCMS deployment (e.g. user identification, search, content access etc.).
8. Deploy and enable custom built CSP module extensions for Drupal to facilitate CSP-based visual assistance in different areas of the website (e.g. search result list and forum representation).

To enable CSP, the website requires two extensions: First, the enabling of CSP-based visual assistance and second is to enable a communication between the website and the CSP Service. The main feature of this extensibility is based on a hooking technique[90] that allows the calling of custom-built functions within the call stack of the WCMS. These hooks are inserted into a module extension and get triggered if the user performs a certain action related to the hook. This allows extensions to hook into the call stack of the WCMS logic (in this case the Drupal Core). Following example illustrates the usage of CSP with the search function of a Drupal based website by using the hooking functionality. In this example the user conducts a search by adding keywords into the search function of a website. A module extension can hook into the call stack of the website and provide CSP based on the user's interaction (Figure 9).

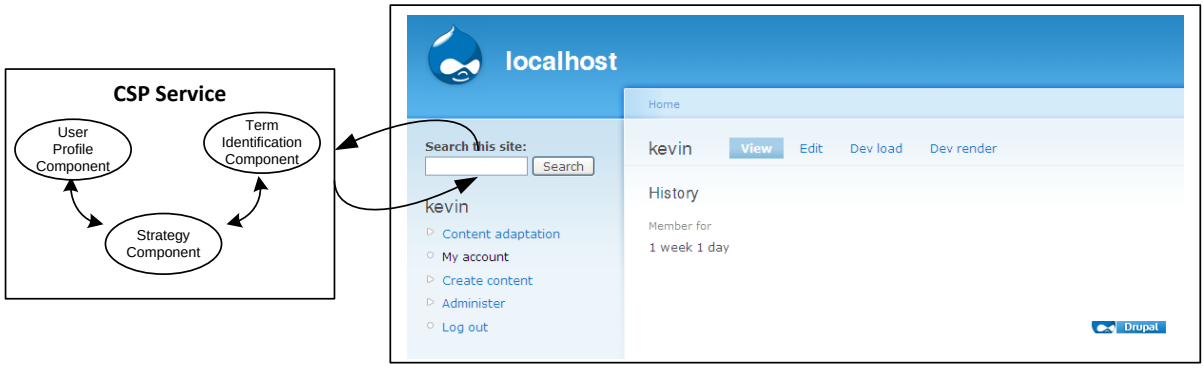


Figure 9 CSP interaction example in relation to search

By using the Drupal hook 'hook_search_preprocess'[91] a custom-built module extension can be triggered once the user executes the search. CSP-based visual assistance can now be

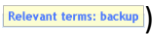
[90] http://api.drupal.org/api/drupal/includes!module.inc/group/hooks/6

[91] http://api.drupal.org/api/drupal/developer!hooks!core.php/function/hook_search_preprocess/6

added to the search result list presentation. In this case the user can chose between search-query related links and results that indicate CSP.

To allow the CSP Service to receive information about the user's content interactions the hook 'hook_node_api'[92] is used. This allows the CSP Service to receive information on what links the user has clicked to allow the user profile component to add terms related to the content.

Following implementation was conducted by using the concepts/modules described above.

1. A custom-made module extension for Drupal was developed that enables hooking into the call stack of the website. The two hooks applied relate to the activity of searching ('hook_search_preprocess') and accessing content ('hook_node_api').
2. The module extension allowed the sending and receiving of terms related to the user activity.
3. The indication of relevancy is based on matching the user's current terms in the user profile with the terms of the content the user has access to.
4. To visualise the CSP relevancy, visual assistance is added to different blocks within the website, such as menu links, search result and forum list. The visual assistance was enabled by influencing the different elements of the WCMS deployment.
5. Links are annotated if the terms of the content behind the links match the user's term profile.
6. The personalisation is displayed as link annotation by adding an icon () beside the relevant links before presentation. All websites used the same image to provide a coherent cross-site experience.
7. To provide information to the user on what terms triggered the adaptive hint a tool tip (

) was provided through mouse over by adding java script and JQuery to the website via a module extension (Figure 10).
8. The link relevancy is based on the unified term model provided by the third party service. The manual hosting WCMS used a highly structured taxonomy model based on the structured manual content.

[92] http://api.drupal.org/api/drupal/developer!hooks!core.php/function/hook_nodeapi/6

9. For model scrutiny both websites are able to display the underlying term model of the user at any time. The assessment of a strategic model was not part of this evaluation.

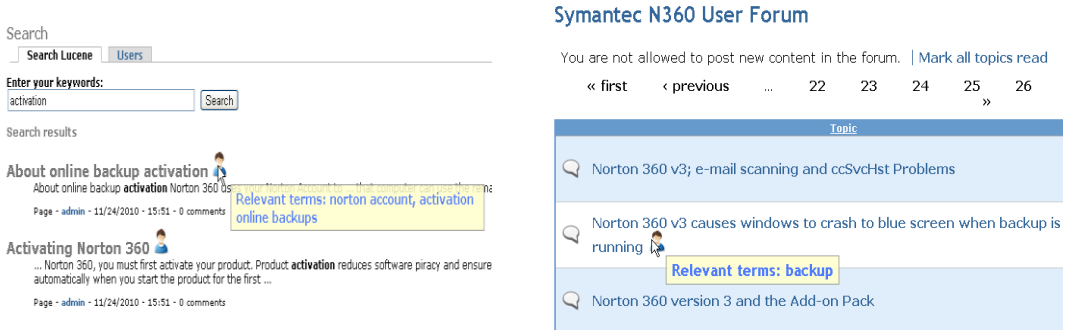


Figure 10 Example of Personalised Link Annotations in Drupal

By using module extensions and link annotations the requirements related to non-intrusive and non-invasive CSP can be fulfilled. Related to the third-party service approach the communication enabled by using a RESTful interface allowing the usage of HTTP commands, such as GET and POST is described in the following section.

The following use case illustrates third party integration to a WCMS for search list annotation based on CSP:

1. User states a query and executes the search.
2. A custom built Drupal plug-in implements the hook 'hook_search_preprocess'.
3. The plug-in requests user information for CSP from the CSP Service.
4. The CSP Service sends the terms of the user back to the website.
5. The CSP extension module within the website hocks into the result list and creates CSP annotations for each result in the result list for which the corresponding content matches the user profile terms.

To summarise note following Figure 11 illustrating the process of CSP integration into Drupal based website. This illustration includes the term identification process outlaid in content preparation section above (4.1.4.1 on page 104). Furthermore this illustrates both cases of already existing websites and setting up a new websites enabling CSP.

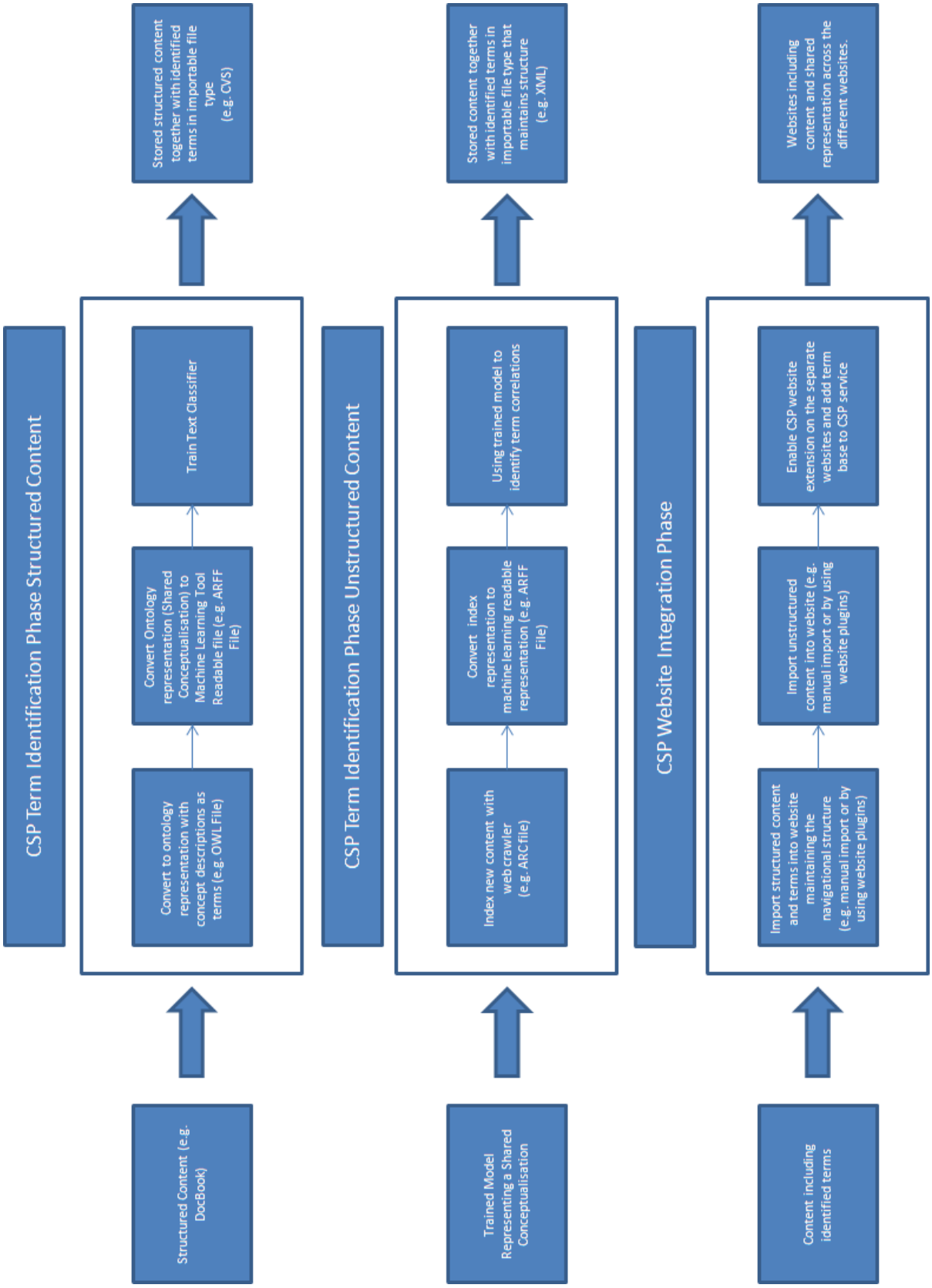


Figure 11 CSP Closed Domain Website Integration Process

4.1.4.6. Communication between the Cross-Site Personalisation Service and WCMSs

A central piece is the CSP Service is to provide a central access point for the different websites. For websites to communicate with the CSP Service and to ensure abstraction (not relying on specific website implementations or subject domains) the communication layer is implemented as RESTful service with the Java based JAX-RS 1.1 (Jersey) library[93]. The service itself (exposing the API) was deployed on a Glassfish application server[94]. The interface between the CSP Service and different websites was facilitated through HTTP GET and HTTP POST. Clients (websites) can use the HTTP GET method to request resources, such as terms related to the previous browsing experience of enable CSP. The HTTP POST method is used to send terms related to the current content viewed by the user back to the CSP Service. Following examples illustrate the simplicity of the approach:

```
PUT /Unite/resources/terms/{username}
Host: myserver
Content-Type: application/xml
<?xml version="1.0"?>
<terms>
  <term>backup</term>
</terms>
```

```
GET /Unite/resources/{username}
Host: myserver
Accept: application/xml
```

The terms sent to the service are added to the user profile via the User Profile Component. For a use case based overview of the communication note Figure 6 in the use case section above.

4.1.5. Process Overview

Following process overview illustrates the dependencies within the approach (Figure 12).

[93] http://jersey.java.net/

[94] http://glassfish.java.net/

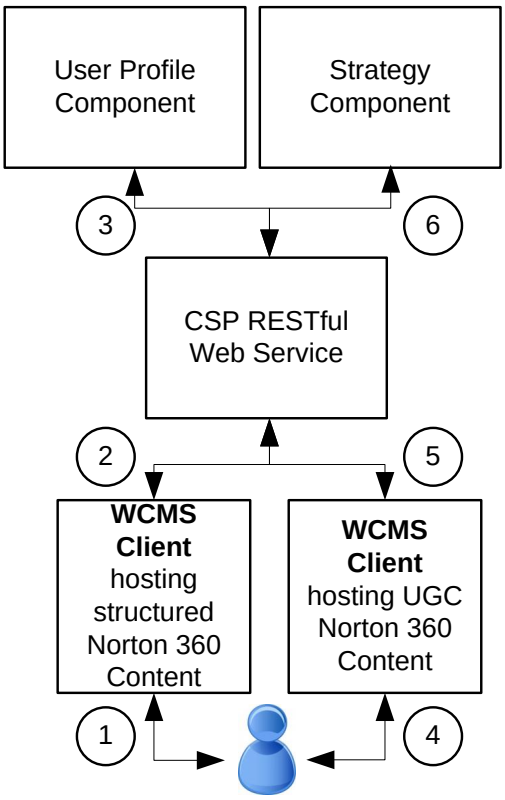


Figure 12 Process Overview Closed Domain Use case

1. User browses to a website hosting structured corporate content. User should be authenticated via sign-in to the website.
2. Website sends CSP Service a user ID. CSP Service uses this ID to ensure the terms (corresponding to the content) are added to the user's profile. Any content the user views triggers the saving of the related terms in the user profile.
3. User Profile Component adds the terms to the user profile.
4. User browses to a second website hosting User Generated Content (UGC). User should be authenticated via sign-in to the website.
5. The webpages requests an informed decision (from the CSP Service together with the user ID) for navigational clues based on the user's cross-site browsing experience.
6. The informed decision is created by the strategy component and based on the user's current user profile and sent back to the websites for CSP-based visual assistance.

4.1.6. Summary

This implementation section describes an implementation related to a use case example applied in a closed domain. For this a specific use case was introduced that is based on the

design requirements introduced in the design chapter of this thesis and that is based on user feedback in an initial user survey (0 page 143). One of the main characteristics of a closed domain is that the content is usually semi-structured. This allows the generation of a trained content model to identify terms in websites (within the domain) that are less structured.

In relation to the use case specific content pre-conditions were discussed. Even though a specific content type (technical documentation) and structure (DocBook) are used, the process is presented to be applicable in different content types and structures. This argument for generality is backed up by the design of the CSP Service for closed domains. The service and the related WCMS module extensions depend on terms representing the content. These terms can either already exist or have to be identified. For identification this section introduces a machine learning based approach that allows the identification of terms based on an initially trained semi-structured content set. It furthermore should be noted that the content used in this use case prototype implementation is based on real-world web content and was consequently used within the evaluation discussed in chapter five of this thesis.

To ensure the implementation addressed the design requirements introduced in the design chapter of this thesis the table overview introduced in the design section related to this use case (Table 3 page 83) is extended. Note the following table to identify the addressing of the design requirements in this use case implementation:

Section	**Concept Description**	**Component** (Figure 5 page)	**Technical Implementation Details**	**Identifier**
3.1.1	A CSP approach should propose assistance that ensures the user is not hindered in their current browsing.	WCMS Module Extensions	Personalised visual assistance is implemented as Link annotation (small icon beside the relevant link). The annotation includes a tooltip	[WUD-1]

			information triggered by a mouse over event indicating the terms that triggered the link annotation.	
3.1.1	A CSP approach should have a unified understanding of the user's browsing space by creating a shared conceptualisation.	Term Identification Component and User Profile Component	For term storage the User Profile Component implements a MySQL database. Within the database the user specific terms are saved.	[WUD-2]
3.1.1	A CSP approach should apply a user modelling technique that does not interrupt the user's browsing. It should therefore be implicit by not requiring the user to explicitly participate in the identification, collection and management of information needs.	WCMS Module Extensions and Term Identification Component	Each click a user makes tigers a Drupal specific hook that informs the CSP module extension about the content accessed. This access is reported to the CSP Service to allow User Profile updating. The terms are sent as parameters together with the user id to the RESTful API	[WUD-3]

			though a PUT command.	
3.1.1	A CSP approach should honour the user's privacy needs by ensuring the different websites the user is browsing do not receive any information about the user cross-site information needs.	Strategy Component	The informed decision is represented as a list of terms. These terms are a result of a matching process between the terms related to links on the current webpage and the terms in the user's profile.	[WUD-4]
3.1.2	A CSP approach should be able to utilise existing content models to create a shared conceptualisation across different websites.	Term Identification Component	Shared Conceptualisation is based on the initial structural analysis of the content. The process includes the creation of a trained content model though the machine learning tool weka. This model is used to identify term correlations in additional non-annotated content.	[WCD-1]

3.1.2	A CSP should be able to use additional services and tools for term identification if necessary.	Term Identification Component	See [WCD-1]	[WCD-2]
3.1.2	Depending on the domain (open or closed) the CSP approach should be able to identify terms of content that is added or updated to website as soon as possible.	WCMS Module extensions and Term Identification Module	Not discussed in this use case.	[WCD-3]
3.1.3	A CSP approach should provide non-intrusive assistance techniques to ensure the user is not hindered in freely browsing.	WCMS Module Extensions	The personalisations are visualised as link annotation by adding a small icon beside CSP relevant links.	[WAD-1]
3.1.3	A CSP approach should affect the website as minimally as possible.	WCMS Module Extensions	A module extension implemented for Drupal based WCMS allows enabling of CSP without interfering with core elements of the website. The interference is based on visual annotation and does not affect the overall websites structure or	[WAD-2]

			design (non-intrusive).	
3.1.3	A CSP approach should not negatively influence core areas of the website (loading time etc.).	WCMS Module Extensions	The information exchanged between the website and the service is kept as minimal as possible (term based) to ensure the websites performance is not influenced.	[WAD-3]
3.1.3	A CSP approach should be based on a design that allows simple integration and interfacing with existing websites.	WCMS Module Extensions	The module extensions can be deployed to the websites and enabled during runtime. However, the pre-extracted terms have to be added to the websites before CSP can be used. Adding terms is not part of the module extension. In this use case a simple term import based on inbuilt taxonomy modules is used.	[WAD-4]
3.1.3	A CSP approach should ensure	Interface Layer	The user can view	[WAD-5]

	that the communication between the CSP approach and the independent websites is flexible and does not depend on the websites technology stack.		the terms in the profile. This implementation displays the terms within the website.	
3.1.3	A CSP approach should ensure device and browser independence.	Interface Layer	See [WAD-5]	[WAD-6]

Table 5 Design Requirements Addressed by a Closed Web Use Environment Use case Prototype Implementation

The following section discusses implementation details related to CSP use cases in open domains.

4.2. Prototype Implementation for Cross-Site Personalisation in Open Domains

This section illustrates the implementation of a prototype to assist users in addressing cross-site information needs within an *open domain*. This section is organised as followed: First the characteristics of open domains are described (4.1.1), which is based on section 3.3.1 of the design chapter. This is followed by introducing a use case example together with use case requirements (4.1.2). The example use case is consistent with the use case used in the evaluation discussed in section 0 of the evaluation chapter. After this content considerations are discussed (4.2.3). The technological architecture and its components are introduced (4.2.4). Finally a process overview (4.2.5) and a summary (4.2.6) are provided.

In the following section open domain characteristics are discussed.

4.2.1. Open Domain Characteristics

Based on the initial use case discussion in the design chapter (3.4) of this thesis following main characteristics of a closed domain can be summarised:

- Users don't need to be authorised to access the content (i.e. open access website therefore the user may not know the website and may not be visiting it on a regular basis).
- The information needs of the user are usually unfocused e.g. exploring/gathering information.
- Open domains often include small-scale websites, forums and blog based website.
- Changes to the websites structure or content are usually applied un-planed and frequent.

4.2.2. Use case Example and Pre-Requisites

This section discusses an example of a use case in an open domain. The specific area chosen for this use case is tourism[95]. In this use case a user browses different websites related to

[95] The content for this use case was kindly provided by the Tourism Board Ireland and is used live on the discoverireland.ie website.

holiday planning.[96] Similar to CSP sue cases in closed domain the websites are independently hosted.

Figure 6 illustrates CSP in an open domain.

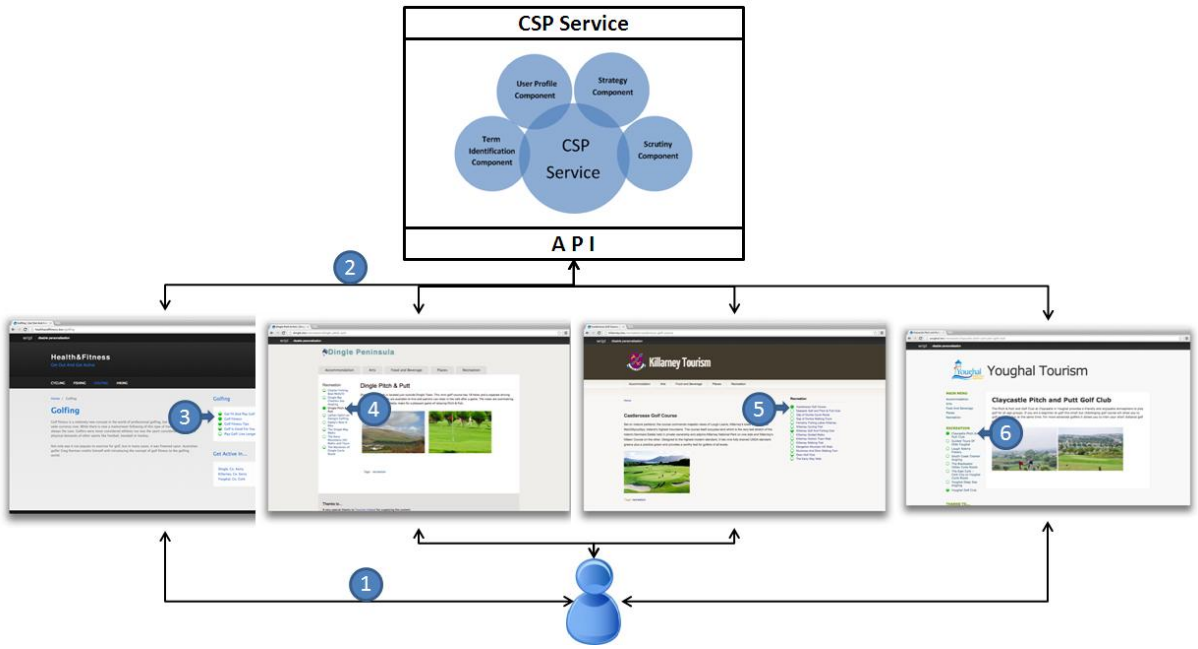


Figure 13 High-level Example CSP Use case in a Open Domain[97]

1. User browses to a website related to an information gathering task in the tourism space.
2. The website requests an informed decision related to the current page (in the case the user has no previous browsing history).
3. The website renders the informed decision as link annotations indicating the relevancy.
4. The user browses to a second website that visualises link annotations based on the user's cross-site information needs (based on the user's browsing behaviour on the first website).
5. The user browses to a third website and receives cross-site adaptive hints.
6. Finally, on a forth website the user receive more relevant link annotations based on the cross-site information needs.

[96] The selection of this use case is based on user feedback provided in the evaluation of the first use case prototype with user indicating usefulness for the CSP approach applied on the open web.

[97] Even though the arrows indicate a one way direction, the CSP approach is implemented to support users browsing in both directions.

Applying CSP to the above illustrated CSP use case in an open domain underlies following pre-requisites:

1. The website should enable communication with the CSP service. This communication should include:
    a. Sending of content for indexing/term identification.
    b. Facilitate CSP-based visual assistance.
    c. Enable/Allow browsing behaviour tracking.
2. Website can use custom build WCMS module extensions and/or access the CSP Service RESTful API directly.
3. The usage of a third-party API for text classification/term identification to ensure fast and reliable term identification.
4. User registration with the CSP Service is required.
5. User sign-in to service via the website is required for CSP enabling.
6. Website should enable/allow third party sign-in/authentication e.g. by using OAuth[98].
7. Website should allow/enable the sending of content updates or added content to the CSP Service for index/term-base update.

The following section discusses specific implementation details based on the above indicated use case pre-requisites. First content considerations are discussed after which the technological architecture is presented.

4.2.3. Content Considerations

In an open domain based use case the content is usually not known at design-time. Furthermore content changes might occur more rapid. Therefore applying CSP to an open domain use case demands a more flexible approach in relation to term identification. The main requirement is to ensure that content updates are considered by the CSP Service at run-time to ensure any informed decision made by the CSP Service is accurate and reflects the current state of the content. To ensure accurate and just in time identification of content related terms a third-party term identification service is used for open domain use cases. For this the approach deploys a pluggable architecture to allow simple

[98] http://tools.ietf.org/html/rfc5849

adding/replacing of external API based term identification services[99]. This furthermore avoids content subject dependencies related to open domain (cross-domain and cross-subject) browsing used cases.

A further consideration is related to content updates. These updates include uploading new content, editing/modifying content, commenting content etc. For this the website should allow the CSP Service to receive updates related to any content base changes.

4.2.4. Technological Architecture and Components

This section describes the technological architecture required for CSP in an open domain. The architecture is based on the high-level architecture introduced in the design chapter of this thesis. In the following the technical architecture (Figure 14) and its components is presented.

To illustrate the specific implementation details technical information is added to the following high-level CSP architecture (Figure 14).

[99] For this the PHP based framework Zend was used.

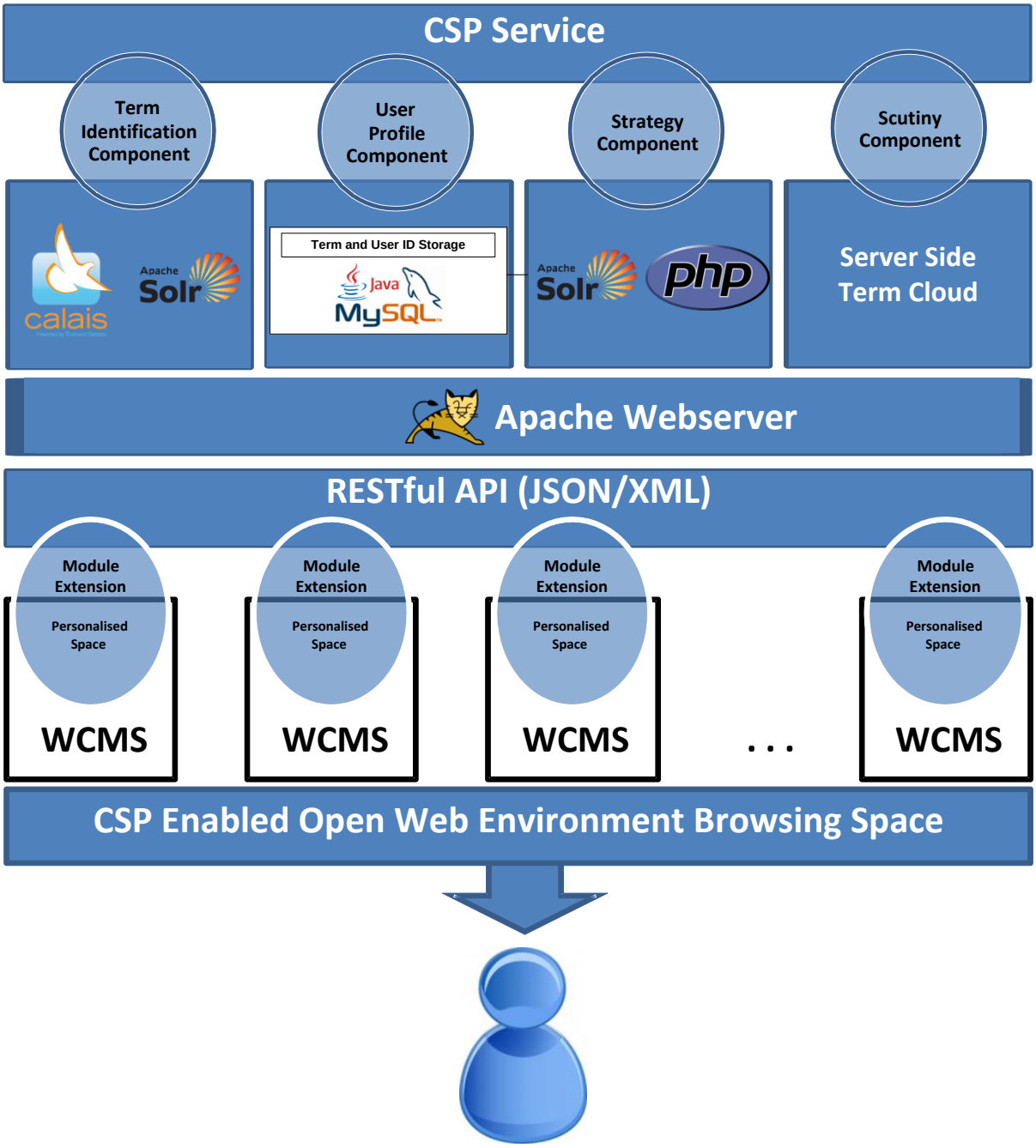


Figure 14 Technical CSP Architecture for Open Domains [100]

Following sections discuss the implementation of the individual components necessary to address this use case with the introduced CSP approach.

[100] The implementation was conducted in PHP/ZEND.

4.2.4.1. Term Identification Component

The term identification component for CSP in open domain should enable term identification as a constant and fast process. Furthermore it should allow the identification of terms across several different subject domains, possibly including different languages and colloquial. Based on this an extendible architecture was implemented. This allows the plugging of additional external term identification services. The current iteration of the CSP approach used in open domain use cases interfaces with OpenCalais[101] due to its wide usage (Rizzo, 2011).[102] An addition investigation was conducted in implementing a solution that provides the same flexibility and maturity as an external term identification service, but is deployable within the CSP Service. OpenNLP[103] was identified as one of the main solutions. However, it requires an extensive training phase and relies on pre-set content to train the service. Based on this shortcoming the Apache Stanbol Project[104] was investigated. It implements a pre-trained OpenNLP instance. However, at the time this research was conducted the project was still in incubation phase.

The result set of OpenCalais includes different types of terms. These include topics (high-level n-grams describing the overall content), social tags (indicating tags that OpenCalais relates to what user 'may' tag the content with), Entities (indicating terms related to high-level entities such as Industry Terms, Local, Position and Technology), finally facts and figures are indicated. After careful consideration and testing the implemented version of CSP for open domain use cases uses the social tags identified by OpenCalais. The testing was conducted with content within the tourism space and focused on how much the identified terms overlap (two-phased evaluation study in the evaluation chapter section 5.4.2 page 168). This overlap is important for CSP to ensure a shared conceptualisation is possible.

The process of term identification is initialised as soon as a website enables CSP (based on this topic by deploying an enabling the CSP extension module). After the website publisher

[101] http://www.opencalais.com/

[102] Prior to selecting OpenCalais different services were investigated with OpenCalais being the most mature and reliable service.

[103] http://opennlp.apache.org/

[104] http://stanbol.apache.org/

enables the CSP module the module sends the content body, title, path id and any associated terms of the openly accessible content to the API of the CSP Service. The service then requests item identification from OpenCalais for the content body sent, which returns social tags associated with a relevancy of either 0.5 or 1. The CSP Service stores the identified terms and the identified relevancy together with the path and website id. The CSP Service therefore implements a table representing all terms identified from the content of the CSP enabled website.

After the terms are identified successfully an index of the website is constructed together with the identified terms. For the index of the website Apache Solr[105] was used. Each website implemented a core and each webpage implementing a Solr document. The identified terms are label as *implicit_terms*. The *solr_document* consist of following custom fields that are added to the Solr core schema file:

- *ID* (Unique identifier of the document)
- *Path* (Path of the content page)
- *Title* (Title of the content page)
- *Body* (Body of the content page)
- *Implicit_terms* (Terms identified by external term identification service OpenCalais)
- *Explicit_term* (Terms sent by the website that are associated to the webpage)

Each core (index) holds a set of documents that can be customised and that represent a page in the index. For CSP each document represents a webpage and each core a website. Depending on the website the user is browsing the CSP Service can identify the appropriate core to generate an informed decision.

For an illustration of the term identification procedure note following process overview:

1. Website enables CSP module extension.
2. Website is registered in CSP Service.
3. CSP Service creates a Solr core for website indexing including a CSP specific schema based on the custom fields listed above.

[105] http://lucene.apache.org/solr/

4. Website sends openly accessible content (body, title and if available associated tags) to the CSP Service via the API Index Endpoint.
5. For term identification the CSP Service sends the body of the content to the external term identification service.
6. A Solr document is created and the populated based on the custom field definitions.
7. The Solr document (representing a webpage) is added to the appropriate Solr core (representing the website).

4.2.4.2. User Profile Component

The user profile is stored in MySQL and stores the terms related to the user's page views and related browsing activities on the page. A user activity is represented by a unique user id, a unique website id, the content path within the website, a list of the inferred terms related to the content path, scroll counter, click counter, touches counter, cut-and-paste counter, key pressed counter and timestamp. The timestamp is important to indicate when the last page access has taken place. This allows the strategy component to ensure more up-to-date activities are prioritised. User activities are sent by the website to the CSP Service via the *Activity_Endpoint* of the API.

For user identification OAuth[106] was implemented. This mechanism allows the website to remain open by allowing the user to sign-in to the CSP Service via the website. Furthermore the approach enables users to authorise the website to receive an informed decision to assist the user.

4.2.4.3. Strategy Component

The strategy component for open domain use cases implements an IR query generation approach. It is implemented based on Apache Solr and uses the existing content index built by the Term Identification Service. Once a website requests an informed decision the strategy component requests the user's current profile terms and the calculated relevancy of the terms. The terms and the related weight are used to create a query on the stored index of the website that requested the informed decision. The result list is filtered based on

[106] http://oauth.net/

the links the website has requested an informed decision for. The result of the filtering is a list containing links and the calculated relevancy (based on the relative position within the result list). The usage of IR was chosen due to its flexibility in relation to term usage and its performance.

The term relevancy calculation is used to assign a weight to the resulting query. The term boost is calculated as follows:

1. Sort all entries in the user profile based on the terms related to the most recent update first.
2. Calculate the relevancy (boost) of each term based on following approximation:
    a. Multiply *CutCopyPaste* actions by 10[107]
    b. Multiply *Clicks* actions by 5
    c. Multiply *Scrolls* actions by 5
    d. Multiply *Touches* actions by 5
    e. Multiply *Keys Pressed* actions by 5
3. Add depreciation factor to calculated boost. The depreciation factor is calculated by dividing the calculated boost by the update position to the power of 2. Therefore the boost of the terms related to the most recent activity has the smallest divider with a denominator of 2. The boost factor is divided by 4, the next by 16 etc. This ensures that the most recent activity has the highest impact within the query.

Each term is now associated with the corresponding boost for query creation. The query follows following structure for each content element within the user profile:

```
$query .= "implicit_tags:(($key)^$value) OR title:(($key)^$value) OR body:(($key)^$value) OR explicit_tags:(($key)^$value) "
```

The resulting query is therefore a concatenation of all terms the user was active on, added with the calculated boost of each term. The query is used on the core of the website the user is currently browsing. The overall result list returned by Solr is based on term matches

[107] CutCopyPaste was assigned with a higher value based on this activity indicating a high relevancy of the content for the user.

within the body field, the title field or the *implicit_tags* field within the Solr documents of the appropriate Solr core.

The search result list is prioritised by search results that relate most to the terms and their boost factor. The boost factor therefore is used to identify the most relevant content items by enabling a higher ranked position in the search result list.

In the next step a matching algorithm identifies the links that are sent by the website within the result list. The link highest on the result list receives the highest relevancy in relation to the amount of links in the list.

It is important to note that the links and their cross-site relevancy, representing an informed decision, are all within the website the user is currently browsing. Note following process overview illustrate the creation of an informed decision in open domain use cases:

1. Website requests an informed decision from the CSP Service by sending a link list to the *Informed_Decision* endpoint of the CSP Services API.
2. CSP Service requests terms and activities from the user profile component.
3. Strategy Component calculates the boost factor of terms based on the above introduced approximation and deprecation factor.
4. Strategy Component generates a query based on the above introduced query syntax.
5. Strategy Component executes the query on the core of the requesting website.
6. Result List is parsed and matched with the initial link list to identify the position of the links within the result list.
7. Based on the position of the links in the result list a score from 1-10 is assigned to the links indicating the relevancy of the link in relation to the user's cross-site profile.
8. The links and the calculated relevancies are added to the list representing the informed decision of the CSP Service.
9. The informed decision is sent back to the website.
10. The website uses the informed decision to visualise the relevancy beside the links.

An alternative strategy, not assessed in this thesis and therefore only mentioned briefly, is the creation of an informed decision without the website sending any links. This alternative strategy allows the generation of recommended links that can be rendered by the website

as a link list or used by the website to indicate relevancy in other website functions, such as website specific result lists. This would allow CSP to be applied similar to Recommender Systems by indicating links related to cross-site information needs. Its applicability would be more intrusive as it would introduce link creation to CSP. To illustrate a CSP strategy that does not involve the request for relevancy of a set of links note following procedure:

1. Website requests an informed decision from the CSP Service via the *Informed_Decision* endpoint.
2. CSP Service requests terms and activities from the user profile component.
3. Strategy Component calculates the boost factor of terms based on the above introduced approximation and deprecation factor.
4. Strategy Component generates a query based on the above introduced query syntax.
5. Strategy Component executes the query on the core of the requesting website.
6. Depending on the amount n of links requested the first n results of the result list would be used. There relevancy would be depending on the position in the list.
7. The links and the calculated relevancies are added to a list. Each link and its cross-site relevancy represent an informed decision.
8. The list is sent back to the website.
9. Website receives a list of links and related relevancy based on the user's cross-site information needs.

Finally, in relation to data privacy requirements it is important to note that no terms are sent to the different websites. The website only receives a list of links with related relevancy indications.

4.2.4.4. Scrutiny Component

The scrutiny component allows the user to see the terms in the user profile together with the indicated relevancy as size. A word cloud for illustration can be used.

4.2.4.5. WCMS Client Integration

To enable a website to offer CSP to their users, a custom build Drupal 7 module extension was deployed. The role of the module was:

1. To enable initial term identification within the CSP Service by sending openly accessible content to the CSP Service.
2. To provide a sign-in mechanism for user to enable CSP.
3. To enable user behaviour tracking once the user has signed in.
4. To provide non-invasive visual assistance once the user is signed in and has authorised the website to receive an informed decision for CSP.
5. To send content updates to the CSP Service on a regular base independent from user being signed in or not.
6. Request an informed decision from the CSP Service.

Drupal was used in this use case implementation based on its wide adoption. Based on the API approach taken by the CSP Service the deployment of extensible modules is not necessary i.e. the website publisher and or the wider open-source community can chose to write addition website extensions/plug-ins. However, using the provided extension significantly decreases development time and based on the extensive research in relation to non-intrusive visual assistance ensures that the user need in not being aversively guided or controlled is addressed.

The WCMS module extension in this use case implementation is focused on a menu structure that is introduced as separate element to a website. The website publisher therefore can replace the standard menu element with the CSP enabled menu element. The menu area looks identical and only is affected if the user enables CSP. The menu structure is tied in with the request of an informed decision. In this use case implementation the influence of the CSP Service is confined to the CSP enabled menu. Therefore the informed decision influences only links within the CSP enabled menu. Prior to rendering the webpage after the user has clicked the CSP module sends the links that the user will see on the resulting webpage to the CSP Service. The CSP Service generates an informed decision that is rendered within the CSP enabled menu.

4.2.4.6. Communication between the Cross-Site Personalisation Service and WCMSs

The interface layer followed the same approach taken as in the CSP Service assisting users in open domain use cases. The main difference is the underlying implementation for open domain use cases is PHP (for closed domain use cases Java was used). The main reason for

this is that PHP is mostly used in open domain related developments. Java more used in closed web enterprise level websites/applications. Following endpoint was used for RESTful Resources:

The communication in this use case implementation is focused on the CSP enabled menu with the CSP. For this following resource endpoints are used:

- INFORMED_DECISSION_ENDPOINT: A website client can post a link list -JSON encoded- requesting an informed decision from the CSP Service endpoint for the link list. The parameter us used to send the link list encoded as JSON.
- INDEX_ENDPOINT: This endpoint enables the sending of content from the website for indexing. The parameters are 'url', 'body' and 'title'.
- ACTIVITY_ENDPOINT: Enables the website to send user activities based on the Drupal hook API.

By using a RESTful API device and website technology independence is ensured and reflecting a key design affordance of user information privacy.

4.2.5. Process Overview

Following process overview illustrates the dependencies within the approach (Figure 12).

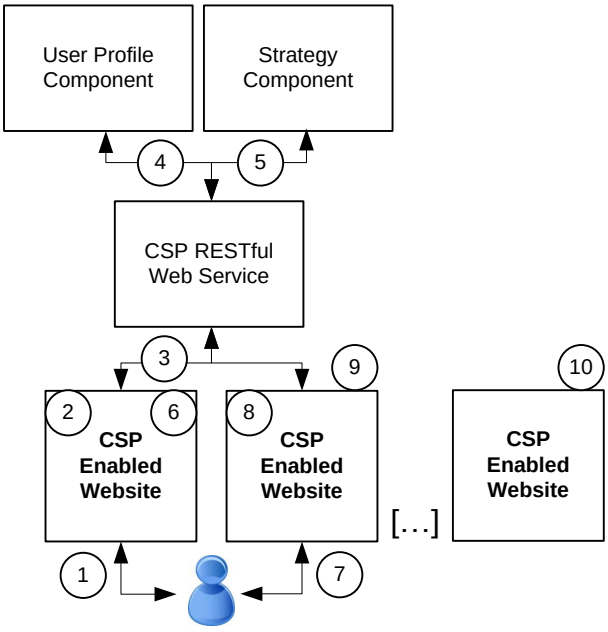


Figure 15 Process Overview for Open Domain Use cases

1. User browses to a CSP enabled website. User can access the website freely.
2. For CSP the user should authorise the website to interface with the service via sign-in authentication. Website sends CSP Service a user ID and authorises the third-third party service to receive and send CSP relevant information (third-party sign-in to enabled CSP not enable website access).
3. Data related to the user's browsing behaviour (e.g. clicks, scrolls, reading time etc.) are continuously sent to the CSP Service (if the user has enabled CSP on the site).
4. The CSP Service updates the user's profile with updated term relevancy.
5. The website requests an informed decision from the CSP Service.
6. The website renders the informed decision in the form of link annotations indicating cross-site link relevancy to the user. As long as the user is on the website and CSP is enabled step 4 to 7 is continuously repeated.
7. The user browses to a second/third and more CSP enabled website.
8. On each website sign-in and authorisation for the CSP Service is required to enable CSP.
9. Steps 4-7 is repeated on each website once the user is signed-in.
10. By adding additional websites the CSP ecosystem can be extended. For this the website should either deploy and enable custom build WCMS modules or facilitate the API.

4.2.6. Summary

This section discussed the implementation of a use case based prototype for CSP open domains. For this specific use case pre-requisites were introduced. The use case pre-requisites were derived from the design requirements introduced in the design chapter of this thesis. One of the main differences between open domains and closed domains is the lacking predictability in relation to the content. This lack is related to both the structure and subject-matter. Therefore, contrary to the first prototype implementation this implementation relies on a content generic term identification method that utilised an external term identification service and Information Retrieval methods.

Even though the implementation introduced in this section is an extension of the first prototype both implementation were connected with the second implementation extending the first.

One extension was in the concept of an informed decision. The first prototype implemented a term matching approach; the second an IR based approach. In the second approach the requesting website sent links for which the CSP service relevancy based on the user's cross-site user profile. Two strategies were introduced, one based on a set list of links and a second based on any link that is within the website.

In relation to the WCMS module extension a single module extension was introduced that implements a menu representation within a Drupal based website. Furthermore a sign-in mechanism based on OAuth was implemented to allow user identification within an open website. A further responsibility of the WCMS module extension is to implement a visual indication of the link relevancy.

To ensure the implementation addressed the design requirements introduced in the design chapter of this thesis the table overview introduced in the design section related to this use case (Table 4 page 93) is extended. The following table illustrates the design requirements addressed by this use case implementation:

Section	**Concept Description**	**Component** (Figure 5 page)	**Technical Implementation Details**	**Identifier**
3.1.1	A CSP approach should propose assistance that ensures the user is not hindered in their current browsing.	WCMS Module Extensions	Non-Intrusive visual guidance through link annotation implemented as Harvey balls[108] with a horizontal relevancy indication. The relevancy is based on a scale from 0-10.	[WUD-1]
3.1.1	A CSP approach should have a	Term Identification	The user profile is	[WUD-2]

[108] http://en.wikipedia.org/wiki/Harvey_Balls

	unified understanding of the user's browsing space by creating a shared conceptualisation.	Component and User Profile Component	implemented as MySQL database. The database stores the page, the related terms and the browsing behaviour of the user (mouse clicks, keyboard clicks and scrolling). Furthermore the difference between the first and the last activity is used to indicate active time also stored in the profile.	
3.1.1	A CSP approach should apply a user modelling technique that does not interrupt the user's browsing. It should therefore be implicit by not requiring the user to explicitly participate in the identification, collection and management of information needs.	WCMS Module Extensions and Term Identification Component	The implicit detection of information needs is based on the browsing behaviour of the user. The content related terms are within the CSP Service and are not shared with the websites.	[WUD-3]
3.1.1	A CSP approach should honour the user's privacy needs by ensuring the different websites	Strategy Component	The informed decision is represented by	[WUD-4]

	the user is browsing do not receive any information about the user cross-site information needs.		links and related cross-site relevancy. No terms are transferred. The website receives no information about how the informed decision was created.	
3.1.2	A CSP approach should be able to utilise existing content models to create a shared conceptualisation across different websites.	Term Identification Component	Shared Conceptualisation is based on the using the external term identification service OpenClais.	[WCD-1]
3.1.2	A CSP should be able to use additional services and tools for term identification if necessary.	Term Identification Component	Not applicable in open domain use cases.	[WCD-2]
3.1.2	Depending on the domain (open or closed) the CSP approach should be able to identify terms of content that is added or updated to website as soon as possible.	WCMS Module extensions and Term Identification Module	The external term identification service OpenClais is used.	[WCD-3]
3.1.3	A CSP approach should provide non-intrusive assistance techniques to ensure the user is not hindered in freely browsing.	WCMS Module Extensions	The personalisations are visualised by the website through WCMS module	[WAD-1]

			extensions. To ensure consistence across the different websites the module extensions implements a menu structure that can be enabled by the website publisher. CSP is confined to this specific structure.	
3.1.3	A CSP approach should affect the website as minimally as possible.	WCMS Module Extensions	WCMS module extension can be deployed at run-time. After enabling the module an indexing procedure is trigger to allow term identification of the websites content.	[WAD-2]
3.1.3	A CSP approach should not negatively influence core areas of the website (loading time etc.).	WCMS Module Extensions	The information exchanged between the website and the service is kept as minimal as	[WAD-3]

			possible (user behaviour information and link lists) to ensure the websites performance is not influenced.	
3.1.3	A CSP approach should be based on a design that allows simple integration and interfacing with existing websites.	WCMS Module Extensions	The module extensions can be deployed to the websites and enabled during runtime. After enabling an indexing procedure should be commenced to allow indexing and term identification within the CSP Service.	[WAD-4]
3.1.3	A CSP approach should ensure that the communication between the CSP approach and the independent websites is flexible and does not depend on the websites technology stack.	Interface Layer	The interface layer is based on a RESTful API to enable communication with different websites.	[WAD-5]
3.1.3	A CSP approach should ensure device and browser independence.	Interface Layer	The generic API provides independent from device and	[WAD-6]

			website specific technologies by only concentrating on user specific information.	

Table 6 Design Requirements Addressed by CSP Prototype Implementation for Open Domains

The following section discusses implementation details related to open domain use cases.

4.3. Conclusion

This chapter introduces two use case based implementations based on the design requirements described in chapter three of this thesis. The first use case implementation focuses on design requirements related to closed domain use cases. The second use case implementation focuses on open domain use cases. Analogue to the use cases, both prototype implementations are separate, but interconnected. To ensure the design requirements are addressed

The development was based on a rapid prototyping approach with the outcome of the first use case implementation influencing the second use case implementation. The main commonality of both prototype implementations is the implementation of a third-party service that interfaces with the websites the user browses across.

Both prototype implementations were implemented to ensure generality in their application in each use case. This argument is based on both implementations designed to be independent from the websites structure, subject and technology stack. For this both prototypes implement a RESTful API and ensure CSP through a term based user profile. In both cases no tight dependency is created between the CSP Service and the websites.

In terms of extensibility of the web browsing space both use cases provide a different approach. The first prototype is built on the characteristics of a closed domain. This domain usually covers similar content subjects and is not extended or updated rapidly. To identify terms within this closed domain a machine learning approach is introduced and based on existing semi-structured content (that usually exists in enterprise based closed domains). This initial term identification process allows the identification of terms of additional content on other (potentially unstructured) content sets on other website throughout the closed domain. Even though the introduced approach requires manual identification of terms in the case new websites are added to the closed domain, it can be argued that the applied approach is sufficient to ensure term identification based CSP is feasible (and meets the design requirements) in closed domains.

The extensibility of open domain follows a different paradigm. For this the implementation introduced is semi-automatic. Websites that are added to the open domain addressed by

the CSP approach are CSP enabled via an automatic Information Retrieval indexing approach in combination with the usage of an external term (name entity) identification service. In addition to simple extensibility this ensures that content changes and updates are reflected rapidly within the CSP Service.

Finally in relation to the connected argument throughout this thesis, that a Web Personalisation techniques that assists user across websites should reflect both the needs and preferences of the website user and of the website publisher by creating a state of equilibrium (introduced as Equilibrium of Web Personalisation). In conclusion it can be argued that by addressing the design requirements derived from the State of the Art review of this thesis (and reflecting the identified gap within the State of the Art) both use case based prototype implementation introduce a applicable approach to assist users in addressing information needs that span independently hosted websites.

The following chapter discusses the findings of a sequence of case based evaluations. The evaluations are based on (and inform) both use case prototype implementations and are conducted mainly as user focused evaluations to assess the appropriateness of both prototype implementations to address the research question, objective and goals of this thesis.

5. Evaluation

The objective of this thesis, derived from the overall research question introduced in chapter one, is to introduce and evaluate a Cross-Site Personalisation (CSP) approach to assist users in addressing information needs that span independently hosted websites. For this the following three goals were introduced: (1) *identify key requirements; (2) design and develop a cross-site Web Personalisation approach and (3) evaluate the appropriateness of the developed approach*. Related to these goals, the following research challenges were stated: (a) *investigate the applicability of CSP in open and closed domain*; (b) *investigate the system requirements to consider the user's privacy concerns*; (c) *investigate the engineering effort on the underlying website to ensure minimal integration effort of Cross-Site Personalisation into websites*.

5.1. Chapter Overview

To address the research objective, goals and challenges of this thesis an iterative design approach was applied and discussed in chapter three. The iterative design interlinked with a series of case study based evaluations and involves the following steps:

1. An initial online survey to investigate the user's cross-site information needs (5.2).
2. A State of the Art review including a separate State of the Art survey of Web-based Content Management Systems (appendix on page 226).
3. The outcome of the survey informed the development of an initial prototype evaluated in the first experiment. This experiment focused on evaluating the appropriateness of applying CSP across websites within closed domains. For this the user performance and the user's self-reported feedback, as well as the technical feasibility of CSP in closed domains was assessed (5.3).
4. Based on the outcome of the first experiment a 2-phased user study was conducted. Its goal was to assess the consistent visual assistance of CSP and the accuracy of both the behavioural analysis of the user's interactions and the resulting personalisations (5.4).
5. The outcome of the 2-phased user study informed a second development iteration that focused on evaluating the appropriateness of CSP within open domains (5.5).

6. A final discussion was added in relation to system related evaluations (5.6).

Figure 16 illustrates the evaluations conducted for this research:

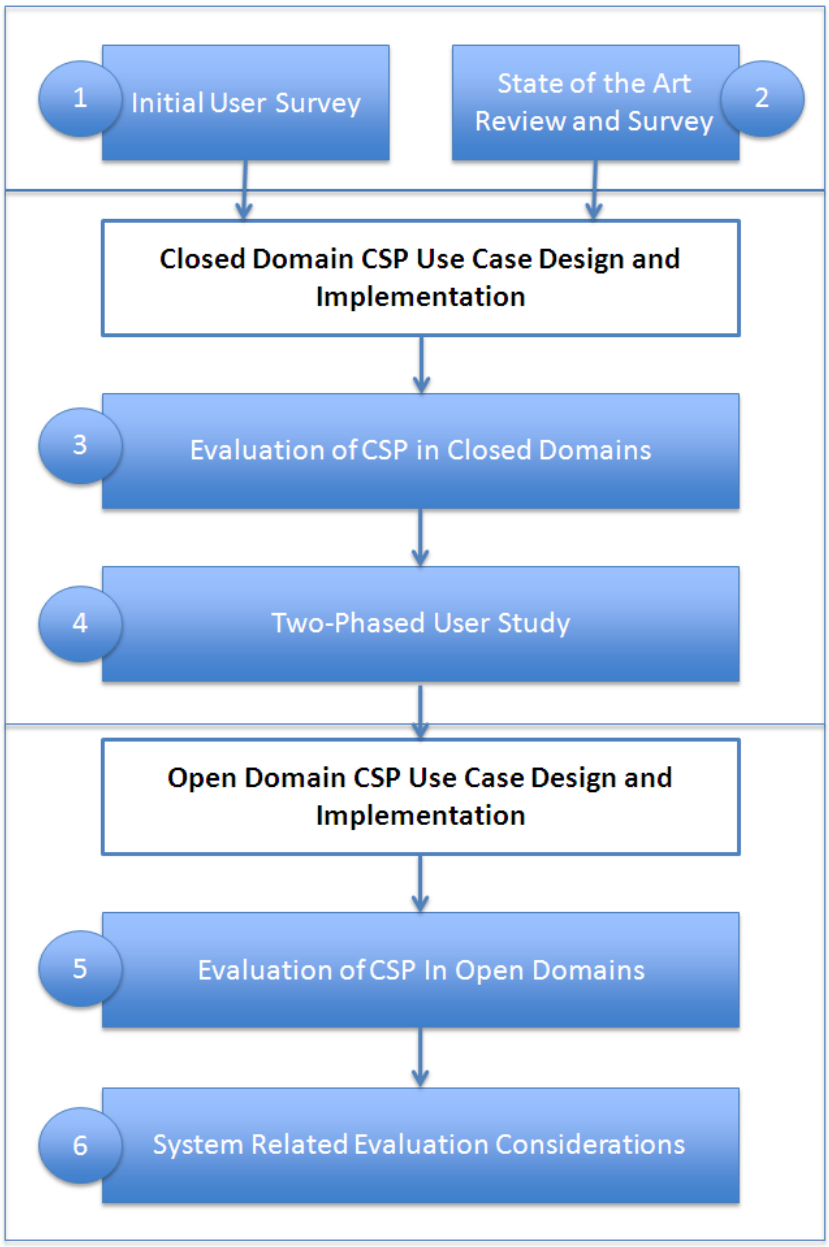


Figure 16 Evaluation Overview

The goal of this chapter is to provide details of the evaluations as well as outcomes and findings. It is structured based on the above mentioned iterative design approach. First, the outcomes of the initial user survey are discussed (5.2). This is followed by a discussion of the findings of the first experiment (5.3). The outcome of the first experiment informed a two-phased user study discussed in connection with the second experiment (5.4). After this a

second experiment is described that focuses on CSP in open domains (5.5). Finally, a discussion about system related findings is added (5.6) and the chapter is concluded with a discussion of the findings (5.7).

5.2. Initial User Survey

5.2.1. Survey Objectives and Setup

An initial survey was conducted to investigate the user's information needs within and across websites. In addition, the types of assistance and user concerns were investigated.

The survey was conducted with 30 participants after an open call for participation through the mailing list of the Knowledge and Data Engineering research group. Most participants were familiar with the research fields of Adaptive Hypermedia (AH) and personalisation. The survey was conducted anonymously and online.[109].

The survey contained 28 questions in total focusing on the identification of cross-site information needs and the identification of concerns related to CSP. The question answer types ranged from simple yes/no answer to open text answers. The following is an excerpt of the findings (for an overview of all 28 questions see appendix on page 243).

5.2.2. Findings

To structure the discussion, the findings are separated into three sections: (1) Identification of the user's information needs, (2) identification of assistance type and (3) identification of concerns.

1. Identification of the user's information needs

To identify the user's information needs more in detail the participants were asked to qualify for what type of information needs Web-based Content Management System (WCMS) are used most. The participants were familiar with the term WCMS that was used in the survey instead of the term website. The focus on WCMS to enable CSP is based on an initial survey conducted in the early stage of this research (appendix on page 226). [110]

To investigation for what WCMS are used the participants were asked to indicate what WCMS are used for most. Based on the results (Figure 17), it can be argued that the type of

[109] For this the survey tool survey monkey was used (http://www.surveymonkey.com/)

[110] It should be noted that recently WCMS extensions are developed that enable the creation of social media related web applications.

information needs addressed by WCMS is related to seeking specific information. The reason for this possibly lies in the domain focus of WCMS. The outcome of this survey indicates that the application of CSP may need to be investigated within a specific domain.

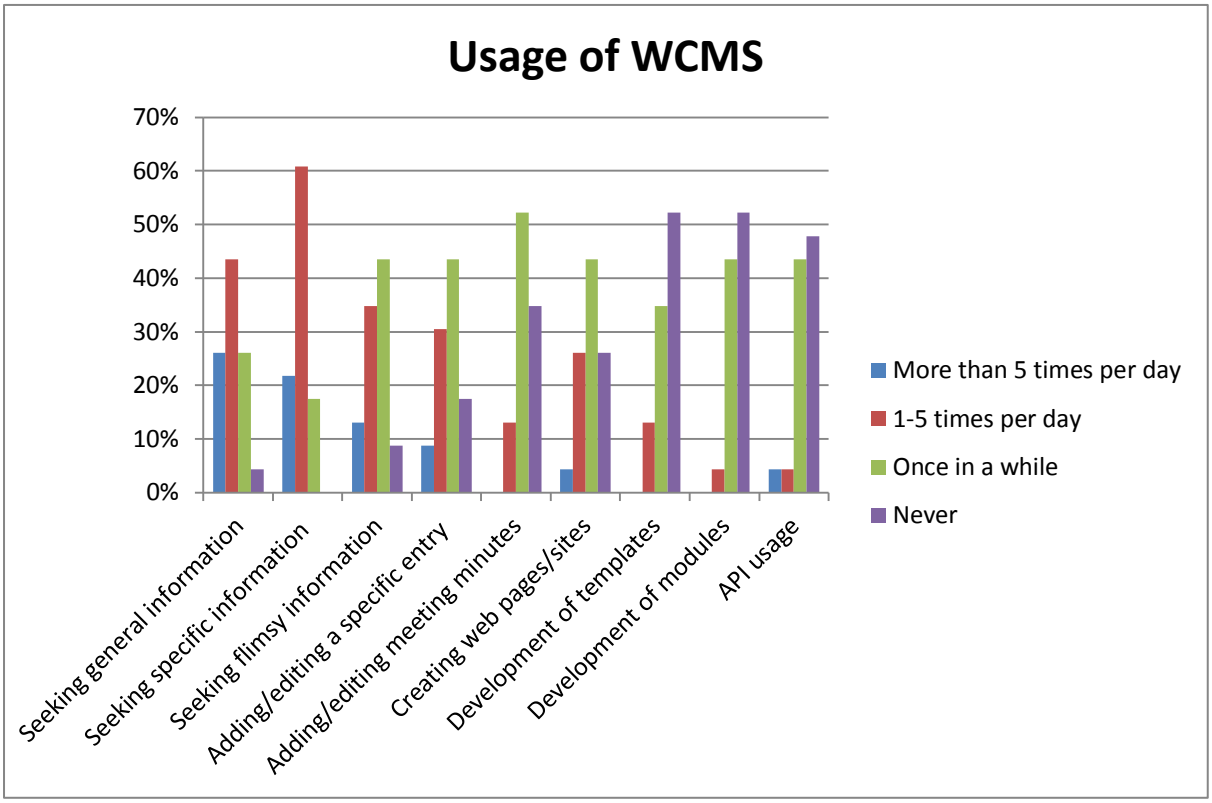


Figure 17 Usages of WCMS

To investigate what type of information representation users prefer, a selection between images, text, video and speech was provided. Interestingly most users indicated text as main preference (Figure 18). This can be seen as indication that CSP applied on text based websites is to be investigated further.

Please specify your content type preference in WCMS implementations?

Answer Options	Very often	Frequently	Sometimes	Occasionally	Rarely	Never	Response Count
If I have a choice I prefer Images	3	7	11	2	0	0	23
If I have a choice I prefer Text	11	9	3	0	0	0	23
If I have a choice I prefer Video	0	1	9	3	8	2	23
If I have a choice I prefer Speech	0	1	3	2	11	6	23
If it depends on the WCMS implementation, please specify							1
						answered question	23
						skipped question	7

Figure 18 Content Type Preferences

In relation to using different WCMS implementations users indicated a stronger bias towards using different WCMS for the same interest/goal/intent (Figure 19). For CSP this result possibly indicates that the integration in one type of WCMS may not be sufficient to provide wide adoption on the open web.

How often do you use different WCMS implementations in seeking the same interest/goal/intent (e.g. Wikipedia and WikiHow)?		
Answer Options	**Response Percent**	**Response Count**
Very often	13.0%	3
Frequently	26.1%	6
Sometimes	21.7%	5
Occasionally	8.7%	2
Rarely	21.7%	5
Never	8.7%	2
	answered question	**23**
	skipped question	**7**

Figure 19 The use of the same WCMS implementation for the same interest/goal/intent

To understand if users experience any type of information overload the participants were asked if they experience any information overload related phenomena (Figure 20). Here, most users did indicate that these phenomena are experienced sometimes.

Do you experience any of the following hypertext/web phenomena when using different WCMS implementation for the same interest/goal/intent?							
Answer Options	**Very often**	**Frequently**	**Sometimes**	**Occasionally**	**Rarely**	**Never**	**Response Count**
Lost in hyperspace (e.g. not sure which WCMS I	2	0	10	5	6	0	23
Cognitive overload (e.g. Too many choices/Too much	1	3	7	7	5	0	23
Development of a mental model of navigation paths	3	5	3	5	4	2	22
						answered question	**23**
						skipped question	**7**

Figure 20 Information Overload Related Phenomena

2. Identification of assistance type

To identify what type of assistance users would prefer, the participants were asked to rate the usefulness of personalisation within and across WCMS (Figure 21).

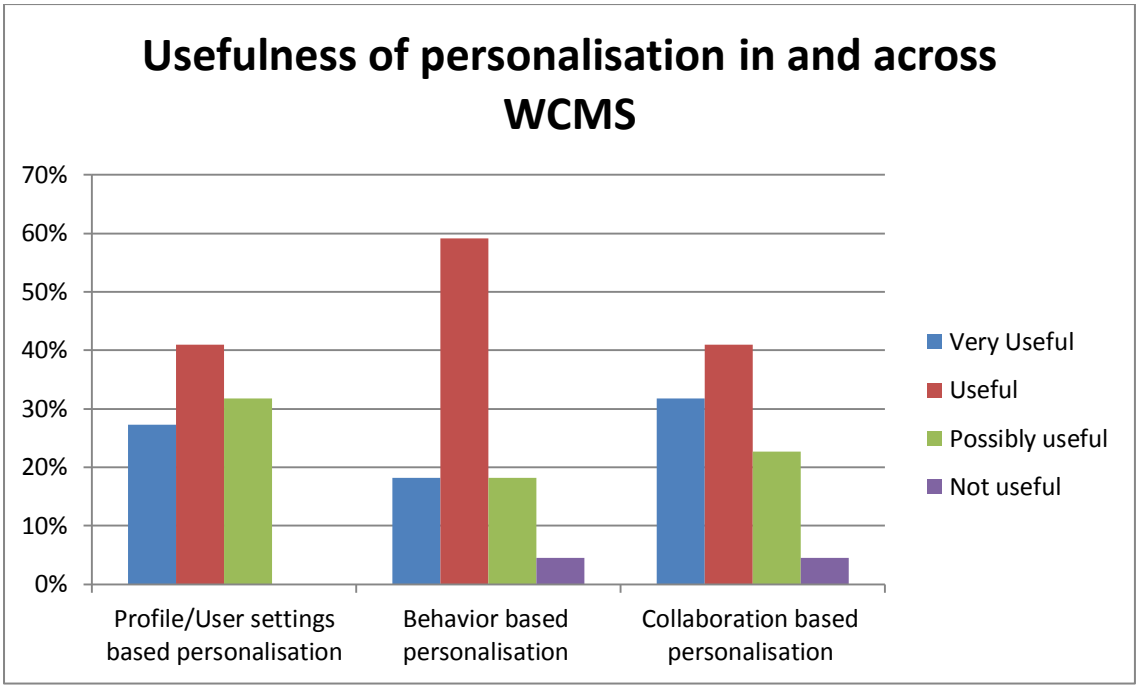


Figure 21 Usefulness of personalisation in and across WCMS

The result indicated that more than half of the participants' rate 'behaviour based personalisation' as useful. This result is interesting in relation to concerns users might have with privacy discussed later in Figure 26 and Figure 27. Here users state a slight discomfort in relation to user profile access. From a user concern point of view it can therefore be argued that the actual collection of user information is not a large concern, but more the storing and access of this information.

The next question asked to the participants was to rate specific types of assistance known in the research areas of Web Personalisation and Adaptive Hypermedia (Figure 22).

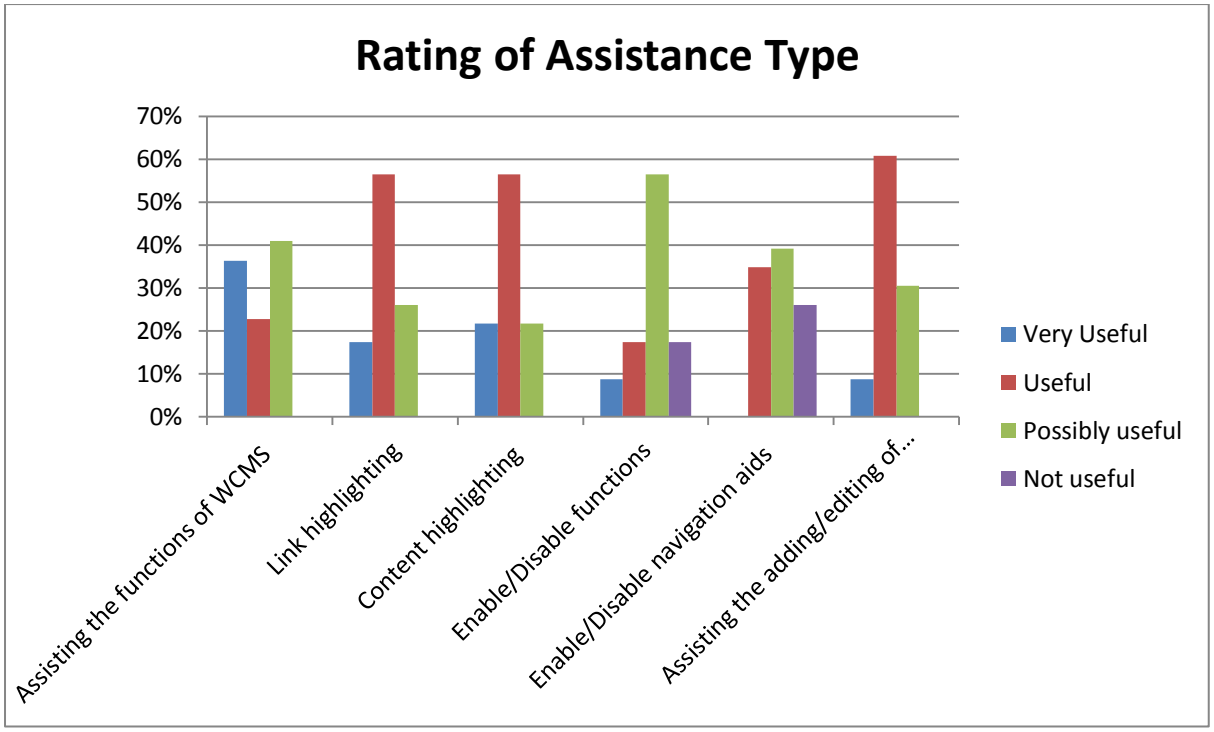


Figure 22 Rating of Assistance Type

Here the participants indicated a slight bias towards link and content highlighting. Interestingly assistance in adding and editing content was also indicated as useful.

To assess if users tend to use assistance tools provided as browser plug-ins, participants were asked if browser plug-ins are used for specific WCMS implementations (e.g. bookmarking tools, password managers etc.). The results (Figure 23) indicate a low usage for plug-ins. It can be argued that based on the low usage of plug-ins users tend to use functionalities, which are integrated in the website and not in the browser as plug-in.

Do you use Browser plug-ins that have an effect on the usage of specific WCMS implementations (E.g. SmarterWiki, Passwordmanagers, dictionaries		
Answer Options	**Response Percent**	**Response Count**
Yes	4.3%	1
No	95.7%	22
If yes, please state the name(s) of the plug-in(s)		0
	answered question	**23**
	skipped question	**7**

Figure 23 Usages of Browser Plug-Ins

3. Identification of concerns

To gauge concerns users may have in relation to an approach that assists users in cross-site information needs, participants were asked questions related to user profile creation, usage and storage (Figure 25 and Figure 24). Surprisingly participants overwhelmingly approved with user information to be created and stored by the WCMS.

Would you allow a WCMS to create a user model based on you browsing behavior (i.e. the WCMS generates it)?		
Answer Options	**Response Percent**	**Response Count**
Yes	90.9%	20
No	9.1%	2
If not sure, please specify		4
	answered question	**22**
	skipped question	**8**

Figure 24 Creation of User Profile based on Browsing Behaviour

Would you allow a WCMS to store and use your user profile (i.e. you have generated it yourself)?		
Answer Options	**Response Percent**	**Response Count**
Yes	86.4%	19
No	13.6%	3
If not sure, please specify		4
	answered question	**22**
	skipped question	**8**

Figure 25 Storing of User Profile in WCMS

Even though the initial response was positive, asked if the participants would allow a third party service to use this model, the response was more reserved. Based on the comments users indicated that the trust would be significantly higher if the third-party service is known and if the user has transparency and control over the information (Figure 26 and Figure 27).

Would you allow a third party service to access your profile/user model in a specific WCMS?		
Answer Options	**Response Percent**	**Response Count**
Yes	43.5%	10
No	56.5%	13
If not sure, please specify		4
	answered question	**23**
	skipped question	**7**

Figure 26 Allowing a third-party Service to Access User Model

Would you allow a third party service to use your profile/user model across different WCMS?		
Answer Options	**Response Percent**	**Response Count**
Yes	47.8%	11
No	52.2%	12
If not sure, please specify		3
	answered question	**23**
	skipped question	**7**

Figure 27 Allowing a third-party Service to use a User Model across different WCMS

It can be concluded that users indicate concerns related to privacy, such as user profile access. However, this concern is almost balanced with a slight improvement if the profile is used across WCMS. For CSP this balanced feedback (11 vs. 12) provides interesting indications that concerns related to user profile access have to be addressed, but do not provide a major barrier.

5.2.3. Discussion

Overall the findings of the initial online survey were encouraging. Related to CSP the following findings are relevant:

- **Content preferences:** Most users indicated to have a preference for text based content. For further CSP studies it can be concluded that text based assistance should be investigated further (Figure 18). This type of assistance therefore would rely on extracting meaning from text based content to assist user in exploring more or similar text based content.
- **CSP usefulness:** Users indicate the usefulness of personalisation within and across WCMS based websites (Figure 21). Behaviour based personalisation was preferred by most users.
- **Assistance type** (Figure 22): In relation to the assistance type link and content highlighting was preferred. For CSP this indicates that the assistance should focus on navigational assistance.
- **Privacy concerns** (Figure 23, Figure 24, Figure 25, Figure 26): Users indicated privacy concern related to user profile usage and access. The results indicate an awareness of users for privacy related matters and therefore should be addressed.

To investigate the technical feasibility of facilitating assistance across websites through a third-party service a State of the Art survey in the extensibility of Web-based Content Management Systems (WCMS) was conducted. The survey indicated the uptake in the usage of WCMS and the possible extensibility towards third-party informed assistance of users.

The next section reflects the findings above and discusses the evaluation of CSP in closed domains. This evaluation is based on the design and implementation of a CSP approach in closed domains as introduced in chapter three and chapter four of this thesis.

5.3. Cross-Site Personalisation in Closed Domains

The goal of this first experiment was to assess the appropriateness of CSP in a closed domain. The focus centres on assessing the user's performance and self-reported feedback based on a comparative study. This section first introduces the experiment objectives (5.3.1). This is followed by a description of the experiment setup (5.3.2) and the baseline comparison (5.3.3). Finally the findings results (5.3.4) and a concluding discussion (5.3.5) are added.

5.3.1. Experiment Objectives

The objectives for this first experiment were derived from the high-level research challenges stated in the introduction of this thesis (1.3):

1. To investigate the applicability of CSP in **closed domains**.
2. To investigate the system requirements to consider the user's **privacy concerns**.
3. To investigate the **engineering effort** on the underlying website to ensure minimal integration effort of Cross-Site Personalisation into websites.

5.3.2. Experimental Setup

The experiment was conducted as a comparative user-study. The study followed a within-subject design and was conducted as an online experiment. The use case was based on real-world customer care data regarding the Symantec Norton 360 product range[111]. Based on the analysis of real-world customer questions[112] a set of tasks was created. Each task consisted of up to six questions ranging from general questions such as *"What are the main features of the Norton 360 backup function?"* to specific questions, such as *"How do I get the Norton toolbar back?"* In each task users were asked to browse within and across two separate WCMS Drupal instances; one hosting structured manual content, the other user generated forum content.

The evaluation was conducted online with 36 volunteering participants from Trinity College Dublin and the University of South Australia. The participants were sourced through email

[111] The content provided by Symantec corresponded to the help section of the Norton 360 product range.

[112] The questions were provided by Symantec.

and forum notifications. The entire experiment was conducted online therefore there was no direct interaction with the participant during the experiment. Furthermore the participants could freely choose when to conduct the experiment. Prior to the experiment each participant received an instructional email illustrating details of the experiment, such as login credentials, length etc. The participants were prompted to log into a website to fill out a consent form, receive further instructions and to conduct a pre-questionnaire to assess the level of knowledge in Norton 360 products, adaptive systems and web research.

The participants were then provided with instructions related to the first system (indicated to the participant as system setting A and consisting of two separate websites). This initial introduction contained a text description and a short video introduction. After the introduction the participant was presented with the first task screen (indicated to the participant as task A). The task screen included a set of questions and text boxes in which the participant could provide the answers. To solve the task the participant was presented with two links. Each link pointing to one WCMS implementation: a website consisting of structured manual content and a website consisting of user generated forum content. The individual links opened in separate browser tabs. To address the task the participant needed to browse across both websites. It was necessary for the participant to use the provided link to ensure the logging of data for the individual WCMS was accurate.

After concluding the task the participant was asked to fill out a System Usability Scale (SUS) questionnaire (Brooke, 1996). The goal of this was to identify any problems with the interface in general, which could potentially distract from the findings. For the participant assessing the CSP enabled websites the SUS questionnaire was followed by a questionnaire specific to the CSP approach. This supplementary questionnaire was based on a 4-point LIKERT scale (1=strongly disagree, 2=disagree, 3=agree, 4=strongly agree). The selection of 4-points was intentional to enforce a preference.

Once the first task was concluded and the questionnaires were filled out the participant was presented with instructions to the second system setting (indicated to the participant as system setting B and consisting of two separate websites). Analogue to the first task, the participant was again presented with instructions and a video. After this the participant was presented with the second task (indicated to the participant as task B). After conducting the

task the participant was asked to fill out another SUS (and if the second system was based on the CSP enabled websites the participant was asked to fill out the CSP specific questionnaire). Finally, the participant was asked to fill out a concluding comparative questionnaire related to the introduced CSP approach.

To avoid a learning effect, each participant was assigned to two out of four tasks. Because this experiment was conducted as within-subject setting each participant conducted one task on each system setting. Therefore each participant conducted one assigned task on the personalisation enabled system setting and one (different) task on the personalisation disabled system settings. The selection of the tasks was determined by the Latin Square design. The order in which the participant evaluated the system was randomised to ensure the CSP enabled and CSP disabled system setting was equally assigned to either the first system setting (system setting A) or the second system setting (system setting B).

The participants were asked to conduct the entire experiment in one session without interruption. During the experiment all system interaction was logged and recorded.

5.3.3. Baseline Comparison

The evaluation was based on comparing two system setting, one setting with CSP enabled and one with CSP disabled. The disabled system setting constituted the baseline comparison. Each system setting consisted of two websites. For both cases the underlying websites were identical in both function and content. In the enabled system setting link annotation ( ) was applied to indicate a cross-site relevant links. Furthermore a tool tip was added displaying the matching terms that triggered the link annotation.

In the website hosting structured manual content, the icon was visible in the menu structure and in the search result list. In the forum hosting website the icon was visible in the topic list of the forum. Figure 28 illustrates an example.

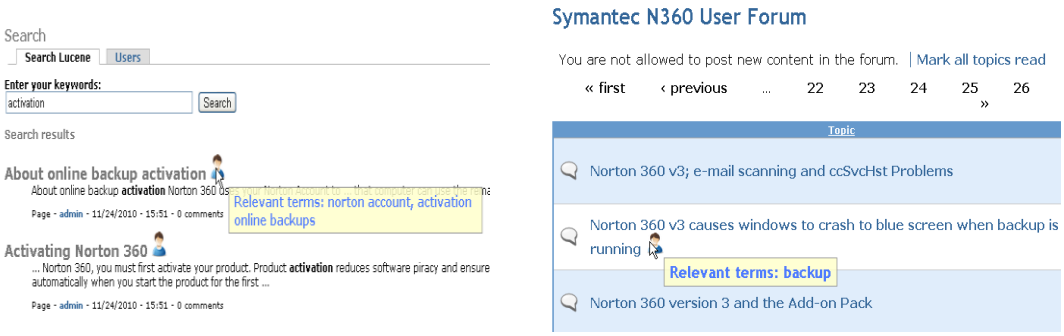


Figure 28 Cross-Site Personalisation indicated as Link Annotation

5.3.4. Results

Overall 36 participants conducted the comparative experiment. The pre-assessment questionnaire indicated that most participants had little to no knowledge about the Norton 360 product range. Furthermore most participants indicated a low level of experience and knowledge in relation to adaptive and personalised approaches. In relation to addressing problems on the web, all participants indicated that they consult forums on a regular base.

In the following, the key findings of both the participants' performance (task time and clicks) and the self-reported feedback of the first experiment are discussed.

User performance findings

Overall each participant conducted two tasks. The *task completion rate* was high and therefore requires no further discussion. The total *task completion time* (without outliers[113]) was calculated for each System setting (Enabled and Disabled).

As illustrated in Figure 29, the overall time of the CSP enabled cohort in Task 1 and Task 4 is higher and in Task 2 and 3 lower as in CSP disabled websites. The overall task completion time of the cohort evaluating the CSP enabled system setting was slightly better with 35512 seconds (SD 433). The CSP disabled cohort had an overall task completion time of 36545 seconds (SD 534).

[113] Outliers were detected by adding (upper outlier boundary) and subtracting (lower outlier boundary) the Mean with 2*Standard Deviation.

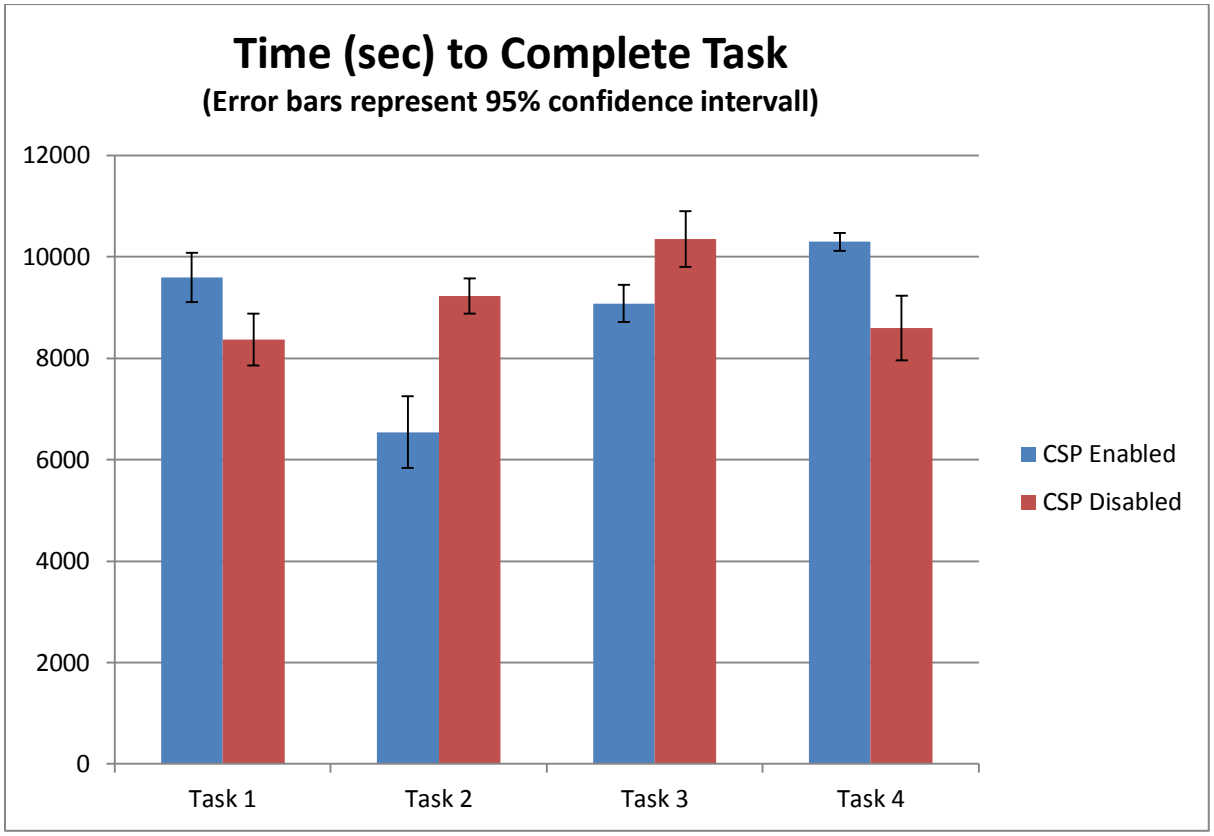


Figure 29 Task Completion Time Closed domain use case Evaluation

To conduct a more in-depth analysis a box chart of the CSP enabled websites and the CSP disabled websites time was created (Figure 30). The overall average completion time in seconds (without outliers) within the CSP enabled websites was lower with 1183 vs. 1352 (median 1071 vs. 1362).

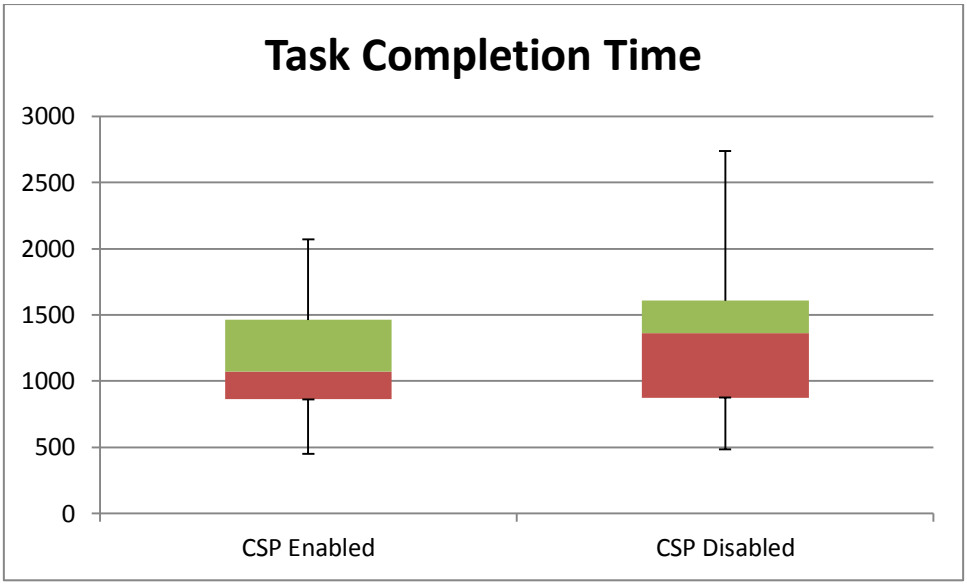


Figure 30 Overall Mean Task Completion Time Closed Domain Use case Evaluation

An analysis of the p-values resulted in a non-significant difference between the task completion times of both system settings (p-value of 0.19)[114]. A possible reason for the high p-value may lie in the fact that the difference between both systems was very subtle with a small visual icon being the only difference. Even though the overall mean comparison is encouraging, a higher number of participants would be required to ensure the mean measurement is significant.

A secondary argument can be made in relation to the nature of the task and the application space of CSP. The task was constructed to assess if CSP can assist users in addressing tasks that require browsing unrelated websites. It can be argued that a task that requires browsing is related to fact-gathering and not fact-finding (found in Information Retrieval related tasks). Therefore the participants might have different ways of assessing content. This individual preference can lead to different reading times not assessed in this experiment.

The analysis of the participants' overall *amount of clicks* is consistent with the analysis of the task completion time indicating a higher rate of content access in the CSP enabled websites (note Figure 31). For this initial experiment any click (within the websites) were counted including repeated access to the same content.

[114] Throughout the discussion of both evaluations all data used in t-tests was checked for normal distribution.

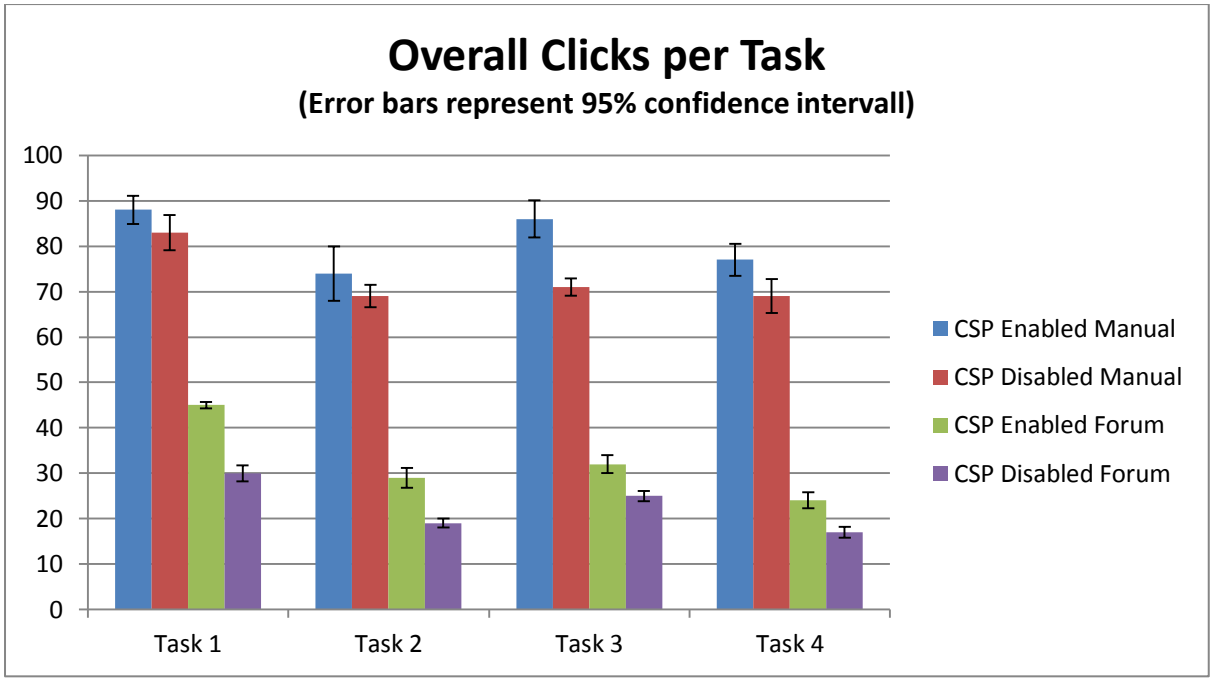


Figure 31 Overall Clicks per Task and Website Closed Domain Use case Evaluation

An analysis of the cumulated clicks per system setting (Figure 32) indicates the same findings. For the CSP Enabled System setting 455 clicks (SD 4.78) were recorded. For the CSP Disabled System setting 383 clicks (SD 4.13)

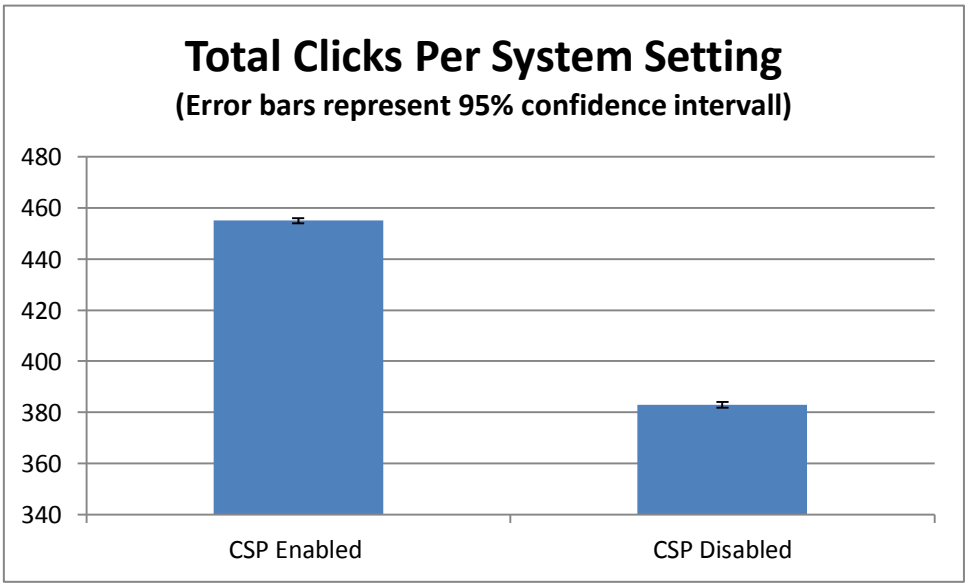


Figure 32 Total Clicks per System Setting Closed Domain Use case Evaluation

Analogue to the time discussion, a box chart was used to gain a more in-depth understanding of the clicks conducted in each system setting (Figure 33). The mean of the

CSP enabled system setting was higher than the mean of the CSP disabled system setting with 15.03 vs. 12.53 clicks with equal median (12 vs. 15).

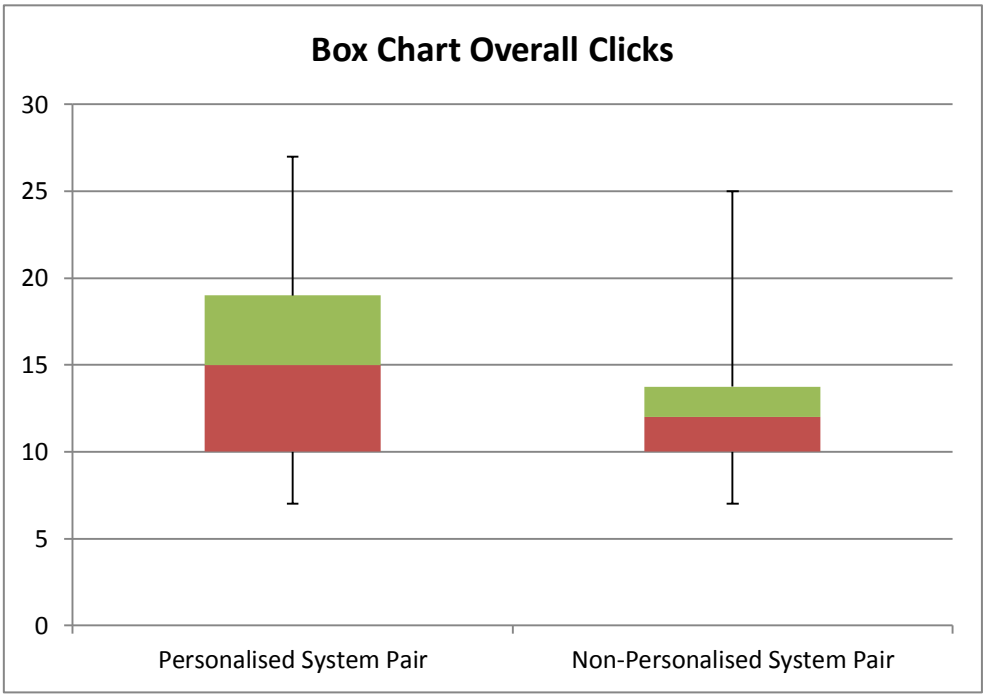


Figure 33 Box Chart Overall Clicks Closed Domain Use case Evaluation

A slightly higher click rate in the CSP enabled cohort can be interpreted as positive due to the users exploring more content. This exploration aspect is important in relation to browsing assistance. However, similar to the user's task completion time, it does highlight critical questions, such as the relation between clicks and time. It can be argued that the increase in clicks may not be interpreted as positive due to the argument that effective guidance should prompt lesser clicks. Both arguments are valid; however, on the contrary it can be argued that by injecting link annotation based on the user cross-site information needs the user becomes more active and motivated to explore additional content related to the task. For example the user might click on links that are not annotated and then decide to also click on the annotated links that would have not been clicked and explored otherwise.

To gain further insight in the impact of CSP on the user's browsing, the following the self-reported findings of the participants are discussed in the following.

Self-Reported Findings

To ensure the overall system experience within the implemented websites is adequate and does not distract the user from the task due to unexpected usability difficulties a System Usability Scale (SUS) was conducted.

The resulting SUS score of both system settings (CSP enabled websites resulted in 68.67 with 31 user responses and the CSP disabled websites in 62.58) indicated that there were no significant usability issues in any one of the two system settings. Even though SUS does not focus specifically on CSP the SUS score provided certainty that the overall website was not designed poorly and therefore did not distract the user with the effect to distort the findings.

To assess the perceived relevance of the CSP approach the participants were asked to indicate the relevance of the adaptive hints when browsing both websites. Most participants agreed (approx. 60%) that the adaptive hints were relevant (Figure 34).

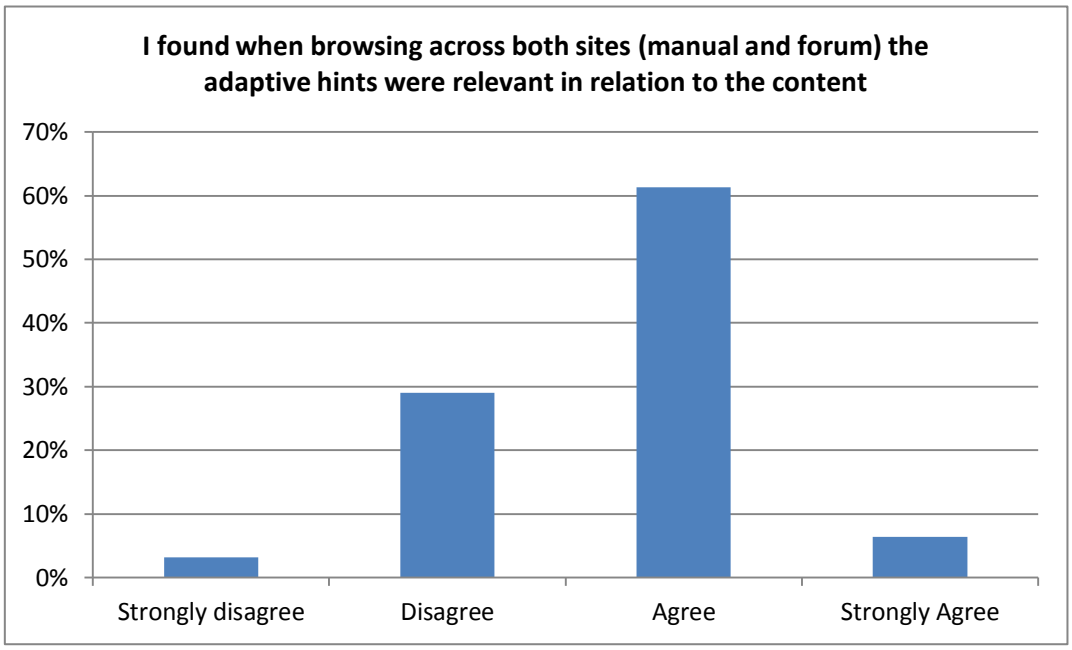


Figure 34 Self-Reported Findings Closed Domain Use case Evaluation Related to Relevancy of Adaptive Hints

Interestingly, asked how useful the cross-site adaptive hints were, participants indicated a mixed reaction with 40% of the participants agreeing and 34% disagreeing (Figure 35).

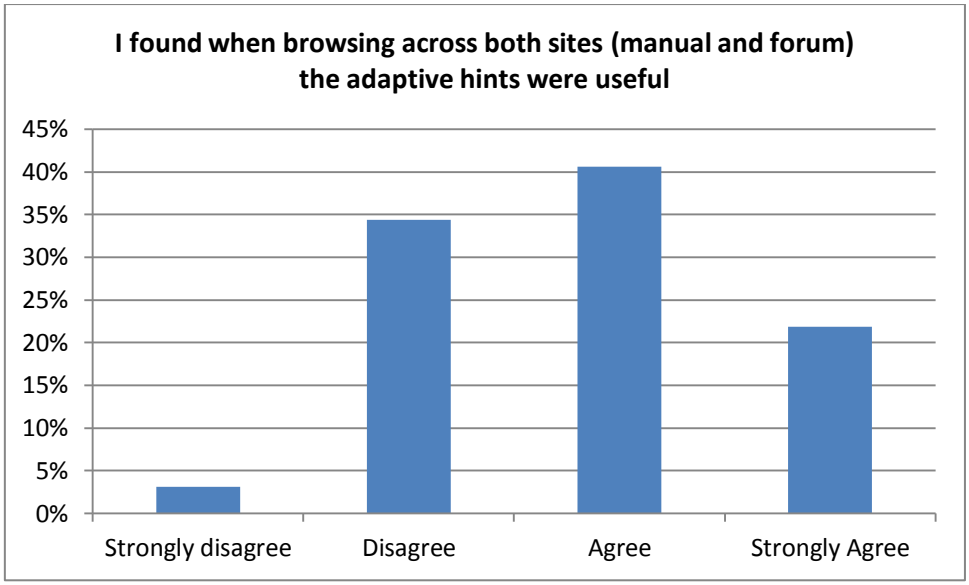


Figure 35 Self-Reported Findings Closed Domain Use case Evaluation: Adaptive Hints In Relation To Adaptive Hint Usefulness

To further examine this mixed feedback, an examination of the participants' perceived awareness was useful. Interestingly more than 60% of users disagreed that they were constantly aware of the adaptive hints on both websites (Figure 36).

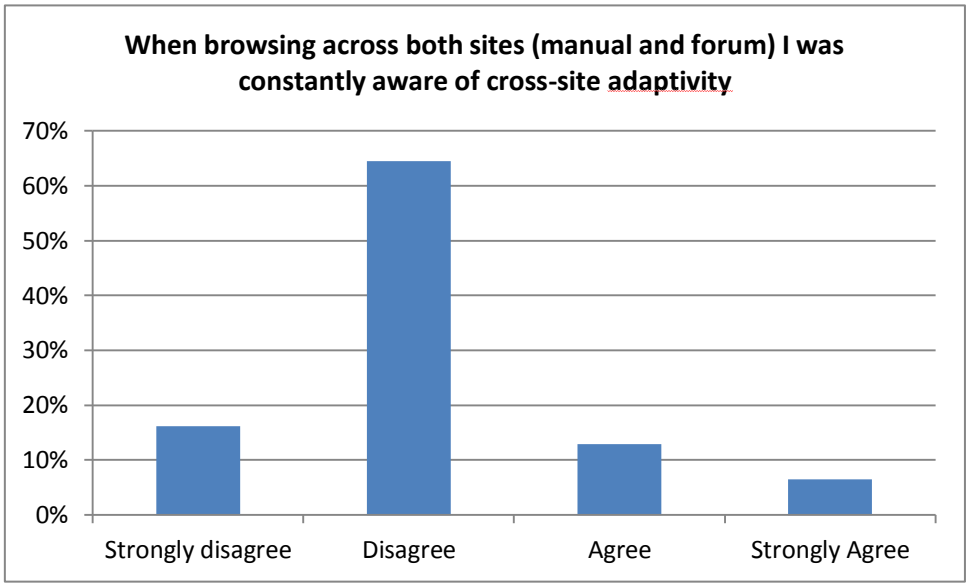


Figure 36 Self-Reported Findings Closed Domain Use case Evaluation: Awareness of Adaptive Hints

Based on this finding, it can be argued that some users may have not been aware of the adaptive hints even though it was clearly described and demonstrated in an introduction video presented to the participant prior to the evaluation. This observation is supported by

comments provided by the participants, such as *"the main problem i had was that i didnt notice the adaptive hints or terms at all really"* or *"the face icon only appeared a couple of times and was always at the page I had already clicked on from the search results"*.

In conclusion it can be stated that the introduced adaptive hint may have been too subtle for some of the participants to notice. Furthermore, with the experiment being conducted online it was not possible to identify if all participants read the introduction text as well as viewed and understood the introduction video.

To investigate CSP further, the participants were asked if CSP would reduce the feeling of being lost. Even though over 30% of the participants disagreed, almost 40% agreed and 30% strongly agreed. This result is encouraging and indicates that the majority of the participants perceived CSP as reducing the feeling of being lost when browsing (Figure 37).

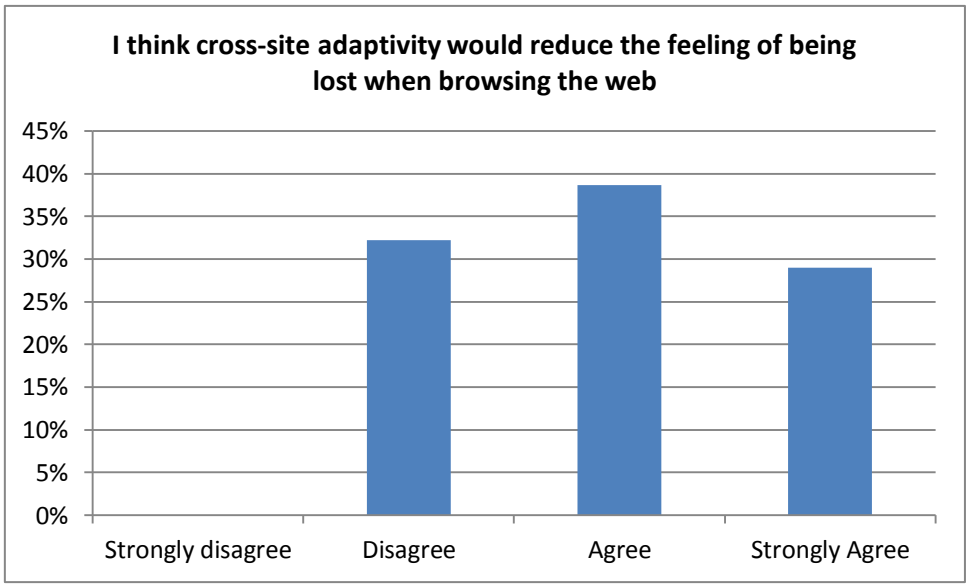


Figure 37 Self-Reported Findings Closed Domain Use case Evaluation: Feeling of being lost

In addition to asking participants questions directly related to the CSP approach supplementary questions were asked.

Asked if users would like to have more control over their user model almost half of the users agreed (40%), 26% strongly agreed and approximately a quarter of the users (26%) disagreed.

In relation to the control aspects almost half of the participants agreed (49%) that they would like to add personal interests to their user model (28% strongly agreed, 19% disagreed and 6% strongly disagreed). Through free text comments some participants stating that the main challenge they see for this CSP is related to privacy. Many participants stated that they would only use the service if it was trusted and if privacy concerns were considered.

Finally, the participants were asked about their overall impression. More than half of the participants agreed that they were satisfied with the performance, assistance and guidance of the CSP enabled websites (19% strongly agreed, 56% agreed, 22% disagreed and 3% strongly disagreed).

This encouraging result was backed up by over 60% of the participants wanting to try the CSP approach in an open web domain (30% strongly agreeing).

To gain more insight into the acceptance of the approach the participants were asked if they would have privacy concerns if used on the open web. The answers were provided in free text with some interesting comments mainly indicating the wish to disable CSP if necessary, and the requirement that the data is to be collected anonymously. Sample answers were "*I would if it could be deactivated. I would prefer the system to be open source to see what it is actually doing*" and "*As long as it collects data anonymously (no personal information just browsing behaviours and interests) and it cannot be trace back to me, it's ok.*"

The participants were also asked what they most liked and disliked about the approach. Users liked the notion of personalisation based on cross-site browsing behaviour. Surprisingly many users indicated that they would welcome more intrusive personalisations.

The main drawback participants indicated was that once they arrived on a webpage and after clicking a recommended link they were left on their own without any further indication on what section of the website is relevant. Other unpopular features were related to the small amount of hints at the start of the browsing process (cold start problem).

5.3.5. Discussion

The following objectives of this first experiment are discussed in this section and were stated as –

1. To investigate the applicability of CSP in **closed domains**.
2. To investigate the system requirements to consider the user's **privacy concerns**.
3. To investigate the **engineering effort** on the underlying website to ensure minimal integration effort of Cross-Site Personalisation into websites.

In relation to the first goal the participants tend to agree that the introduced CSP approach was helpful and relevant to assist browsing across two separate websites. Furthermore the participants tended to complete the tasks faster by using the CSP approach (Figure 29 on page 155). Interestingly a higher click rate was found in the CSP enabled cohort.

In relation to the second objective the feedback of the participants provided valuable insights in concerns and needs related to the introduced CSP approach. The main concerns identified were related to privacy concerns, user model control/scrutiny and the usage of CSP on the open web.

The third objective was addressed by providing WCMS module plug-ins that facilitates both the communication with the CSP service and the link indications in the website. The main effort in relation to integration was related to term identification.

An overarching discovery was that users indicated that the introduced CSP approach was too non-intrusive. This observation is backed by the subtle difference in the user performance between the two cohorts (CSP enabled websites/CSP disabled websites). This argument is supported by the participants' self-reported responses and comments. Even though the participant's viewed a video prior to the experiment explaining the introduced CSP approach, some participants commented that they were not aware of any difference between the two settings.

The outcome of the first experiment that focused on evaluating the initial CSP prototype in a closed domain was encouraging. The main reason for encouragement was that it was possible to provide a CSP Service to communicate with two separate websites and provide

consistent user assistance. Further reasons for encouragement were: (1) Users indicated the relevancy of the provided personalisations (Figure 34); (2) A measurable difference in the performance of both systems pairs was recorded (Figure 29); (3) Users indicated strong interest in using the approach in open domains. However, less encouraging observations where: (i) Users perceived the introduced personalisations as too subtle (Figure 36); (ii) Users indicated privacy concerns; (iii) Task design and experiment setup requires more concentration on exploration and may require lab based assessment to allow open questioning of the user after completing the task.

Finally, this first experiment indicated the need for an even more in-depth discussion in relation to examining the usefulness of Web Personalisation within and across websites. The experiment was conducted online; therefore the examiner had no chance to discuss details of the experiment with the participant after it was concluded. Furthermore, many tasks were addressed by both searching and exploring. It is tempting to conduct search tasks in web experiments due to the straightforward measurement of performance and error. In exploration based tasks, such as browsing, which are of high importance for this research, it is difficult to assess the usefulness, effectiveness and efficiency through an anonymous and remote online experiment. One example of this difficulty is that it is not clear if the user followed a link because of the adaptive highlighting of it, or because the title of the link matched the user's expectations. To gain full insight the experiment would require the usage of more intrusive observation methods, such as eye tracking and/or video surveillance (Hauger et al., 2011). However, intrusive observations interfere with the natural open-web browsing behaviour of the user. Therefore a balance between more intrusive observation and the participants' natural browsing environment should be found. These evaluation challenges are not new to this research field and have been highlighted throughout different publications (Bernstein et al., 2011; Kohavi et al., 2012; Paramythis et al., 2010).

5.4. Two-Phased User Study: CSP-based Visual Assistance and Relevance

To gain more in-depth understanding of the individual elements within the CSP approach a 2-phased user study prior to a second comparative experiment was conducted. This approach was inspired by the layered evolution introduced by Paramythis et al. (2010). The first phase addressed link-based visual assistance challenges of CSP, the second focused on the accuracy of the behaviour-based link hints.

Overall 10 participants volunteered for both phases of the study. The participants were all members of the Knowledge and Data Engineering Group of Trinity College Dublin and consisted of full time PhD students and academic members of staff. All studies were conducted in a lab environment. The feedback was collected through questionnaires and open interviews.[115]

5.4.1. Phase 1: Investigation of Link-based Visual Assistance

To investigate link-based visual assistance for CSP, a set of image based design studies were conducted. The participants were asked to provide feedback on each link-based visual assistance type via reaction cards[116] and questionnaires. The reaction card approach, proposed by Benedek and Miner at Microsoft, consists of 118 words that allow the user to react to the design presented.

For each visual assistance type the user was presented with a set of real-world webpage examples. These examples were manually extended with the appropriate link-based visual assistance.

Four different link-based visual assistance types were assessed: Link highlighting (Figure 38); Link annotation with a star symbol (Figure 39); Link annotation with a circular symbol (Figure 40) and a small dotted bar (Figure 41).

[115] The following discussion is an excerpt of the full report see appendix on page 250.

[116] http://www.microsoft.com/usability/uepostings/desirabilitytoolkit.doc

Figure 38 Link Highlighting

Figure 39 Link Annotation with a star symbol

Figure 40 Link Annotation with a circular symbol

Figure 41 Small Dotted Bar

For each link-based visual assistance the indication of relevancy was assessed. In link highlighting the relevance was indicated based on colour intensity. In link annotation with star symbols the colour scheme bronze, silver and gold was applied (gold as most relevant). For link annotation with a circular symbol Harvey balls[117] were chosen. Finally, for the dotted bar relevancy was indicated by filling three separate sections. Each participant was presented with at least three different website examples for each link-based visual assistance. This was to avoid bias towards a specific website or layout. In the following the outcome of the study is presented.

Findings Link Highlighting

The overall reaction of the participants was positive. Most indicated link highlighting works well if not too many elements are highlighted. However, the relevancy based on colour intensity was indicated as not obvious enough. For this users indicated that a numbering might be useful. The reaction card feedback was mostly positive with 32 positive words and 10 negative words. More than 50% of the users agreed that link highlighting would influence

[117] http://en.wikipedia.org/wiki/Harvey_Balls

their browsing behaviour in a positive way and that the highlighting was not distracting. Users indicated that they preferred that the strategy was simple, clean and time efficient. Users disliked that it was not obvious and sometimes hard to see. As an improvement, some users mentioned they would like to be able to select the colour scheme.

Findings Link Annotation with a Start Icon

The star icon resulted in the selection of 37 positive and 2 negative word reactions. Even though the initial reaction was positive, most users indicated that the link-based visual assistance was not clear and not intuitive. Most users agreed that the annotation was noticeable, intuitive and that it was not perceived as confusing. Furthermore almost all users agreed that the strategy was less intrusive and that it would affect browsing in a positive way.

Findings Link Annotation with a Circular Symbol

The reaction of the users for the circular symbol was positive with 46 positive word selections and 0 negative word selections. Most comments were made in relation to how the circle is filled. The filling used reminded some users of a clock that indicated time and not relevancy.

Users liked most that the colour was noticeable and that it contained a lot of information in a small space. Surprisingly some users mentioned that it felt personal. To improve the strategy, participants indicated that the circle should fill horizontally depending on the level of relevancy.

Findings Link Annotation with Bar (dots)

This strategy was perceived as mostly negative with only 5 positive words selected and 19 negative. Participants liked that it was clear and simple. However, most comments were negative, such as that the icon was too small and it was hard to identify the relevancy.

Conclusion of Findings

Based on the positive feedback and reaction the link annotation circle was selected for the final experiment. Furthermore, the relevancy filling was change from quarters to horizontal

filling to avoid the feeling of time relevancy. To ensure the decisions meet the expectations of the participants the final link-based visual assistance was presented to the users, the reaction was mostly positive with users selecting 57 positive words and only 2 negative words.

5.4.2. Phase 2: Behavioural Tracking and Recommendation Creation

The main focus of this second phase of the user study was to investigate the accuracy of the link relevancy indicated by the CSP Service. This investigating was conducted in two steps. The goal of the first step was to identify if the terms and the related relevancies that are inferred by the system are accurate. The second step investigated if the link relevancy calculated by the system, based on the terms and related relevancy, is accurate. Both steps were conducted with the same participants and same laboratory setting as in the first phase of the user study.

For the first step the participant was asked to read content related to health and fitness activities on one website. The website used was setup up prior to the experiment and included 5 links. The content was pre-analysed for term extraction. Furthermore the users browsing behaviour (clicks, scrolls and keyword presses) was used to assess the relevancy of the content related terms. After the task each participant was asked to write down five terms that relate best to the task. Then the participants were asked to rate each term between 1-10 depending on its relevancy. After this, each participant was presented with a term cloud that indicated the terms and the relevancy the system had inferred based on the users browsing behaviour. The users were asked to indicate if the terms represented their browsing behaviour.

In a second step each participant was asked to browse to a second website. On this website ten links were presented beside the content. The participants were asked to rate each link (0 to 10) in relation to their cross-site browsing interest. The same test was conducted on two other websites.

5.4.3. Discussion

In relation to the first step users were asked to indicate if the system inferred word cloud matches the user's expectations. Overall 3 strongly agreed, 4 agreed, 3 neutral and 2

participants disagreed. Asked if the term size, that indicated the level of relevancy, was accurate, 2 participants disagreed, 3 here neutral, 4 agreed and 2 strongly agreed. It should be noted that the term model was slightly different for each participant based on the different browsing behaviour.

In the second step the participant's indicated link relevancy was compared with the relevancy the system calculated based on the users profile. In most cases the relevancy overlapped. However, in some cases the relevancy calculation of the system tended to be less extreme at the edges of the scale. For example the participants often rated links as either 0 or 10 and the system often rated a relevant link as 7 or 8 and a lesser relevant link as 2 or 3.

It can be concluded that the difference between the user's judgement and the systems calculations (based on the users' behaviour) were accurate enough to proceed with a second experiment.

5.5. Cross-Site Personalisation in Open Domains

This second evaluation was based on the outcome of the first evaluation and the 2-phase design study. The goal was to extend the initial prototype and to assess how appropriately the introduced CSP approach assists users in open domains. This section is structured analogue to the first experiment.

5.5.1. Experiment Objectives

The second experiment aims to extend the findings of the first experiment through self-reported feedback and the analysis of user performance metrics, such as time on task and clicks. The following objectives were derived from the high-level research challenges stated in the introduction of this thesis (1.3):

1. To investigate the applicability of Cross-Site Personalisation in **open domains**.
2. To investigate the system requirements to consider the user's **privacy concerns**.
3. To investigate the **engineering effort** on the underlying website to ensure minimal integration effort of Cross-Site Personalisation into websites.

5.5.2. Experimental Setup

The second experiment focused on an open domain use case in the tourism space. The content was provided by Failte Ireland and is used live[118]. The experiment consisted of three websites. The first website acted as entry website and was based on health and fitness related content in the four activities Golfing, Hiking, Walking and Cycling. The purpose of the entry website was to train the users profile to enable CSP across the remaining two websites. The two additional websites were related to tourism and were deployed each relating to different locations (Killarney and Dingle). Each website was hosted as a separate WCMS implementation with different templates to ensure a different look and feel on each website. To evaluate the effect of CSP on the user's performance and to assess the self-reported feedback it was conducted as a between-subject experiment, with each cohort consisting of equal numbers. Each participant was randomly assigned to the CSP enabled or CSP disabled group. Similar to the first experiment two different system settings were used.

[118] www.discoverireland.ie

The first system setting consisted of all four websites with CSP enabled the second with all four websites CSP disabled. The underlying websites including the layout and content were equal in both system settings.

Overall 60 participants took part in the user-focused study. The participants were sourced through school mailing lists and black board notices throughout Trinity College Dublin. The experiment was conducted in a closed lab environment to reduce distraction, and allowing further feedback through an unstructured interview. The evaluation data was collected through questionnaires, system logging and unstructured interviews.

The duration of each evaluation was approximately one hour and the overall evaluation including the two-phased design study was conducted over a timeframe of six months. One of the goals of the evaluation was to ensure the experiment setup was as natural as possible. To ensure a realistic browsing environment the tasks were based on real-world fact gathering tasks and the experiment environment was set up without any intrusive recording devices.

Each participant was asked to conduct one browsing tasks related to planning a holiday around the activity of Golf. The goal of the experiment was to explore and collect information that would allow the participant to convince friends that a holiday around the activity Golf is worth pursuing. After filling out the ethics form and informing the participant what information will be recorded, a short description of the task was presented. The participant was asked to provide all answers in handwritten form to ensure the participants were able to answer the task questions without interrupting the browsing setup (which would have been necessary in an online form). Finally, the participant was given a chance to ask questions. Once the participant was ready to start, task time was recorded and the examiner left the room. The participant was asked to stop the clock once the task was completed and then call the examiner back into the room. Different than in the first experiment the participants were not informed about CSP prior to the experiment. The reason for this was to ensure the neutral assessment of the CSP approach.

After the task the participants were asked to complete an online survey with 22 questions. The cohort with the personalised setting was asked to fill out a second short paper

questionnaire with the examiner in the room in order to trigger open questions and responses in relation to the introduced CSP approach. For this the participant was made aware of the introduced approach.

5.5.3. Baseline Comparison

The experiment was conducted comparatively as a between-subject design. Each cohort was presented with a total of three independently hosted websites. The first website was used to train the participant's profile. The latter two evaluated the appropriateness of the CSP approach. All three websites were hosting real-world content in the area of recreation and tourism. The baseline system was deployed without any personalisation features enabled. From the participants perspective there was a difference how the side menu presented links. The CSP enabled websites implemented a green link annotation icon beside every link in the menu. Each annotation icon was filled depending on the level of relevance the link had. All other website features were identical. See Figure 42 illustrating the added CSP based annotations.

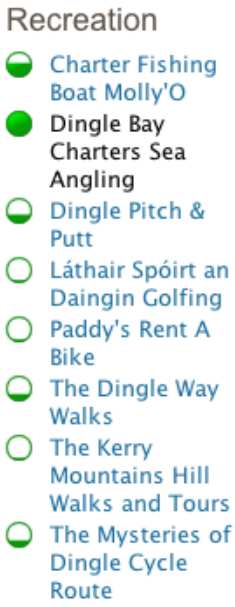


Figure 42 Example of CSP Based Annotation on a website

The following discussion of the findings is split into user performance and self-reported feedback.

5.5.4. Results User Performance

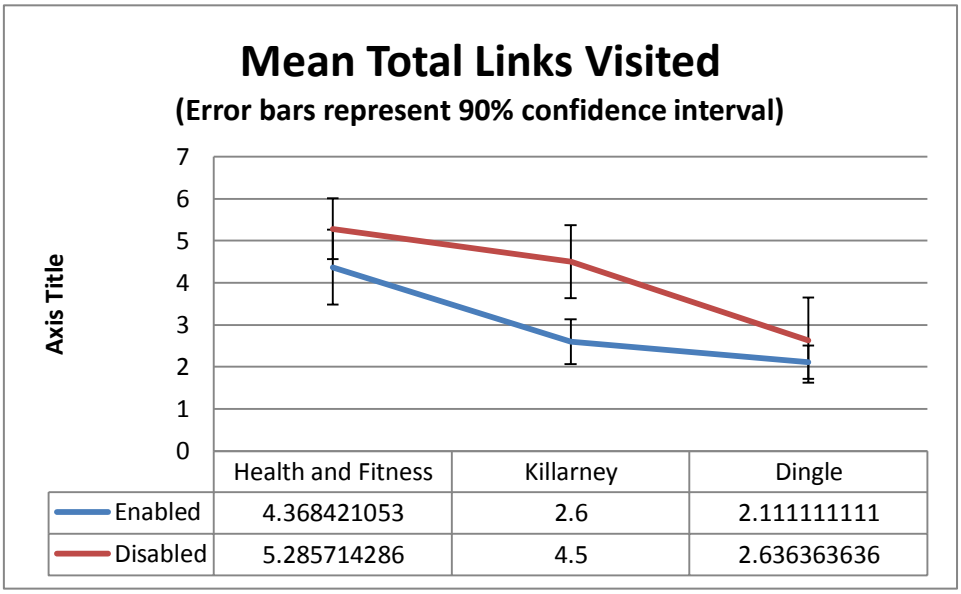
The findings related to the user's performance were divided into click-related metrics and time-related.

Click-related Findings

To discuss the influence of the CSP approach on the click behaviour of the user, the overall clicks were analysed.

The comparison of the total clicks (without outliers and including repeated page views) within the cohort assessing the CSP enabled websites was 173 vs. 268 in the CSP disabled websites. The individual results were 83 (enabled) vs. 111 (disabled) clicks in the first 'Health and Fitness' website, 52 (enabled) vs. 99 (disabled) in the Killarney website and 38 (enabled) vs. 58 clicks (disabled) in the final website 'Dingle'.

To further investigate this encouraging result the average clicks of each individual website were compared (Figure 43).



	Health and Fitness	Killarney	Dingle
Enabled	4.368421053	2.6	2.111111111
Disabled	5.285714286	4.5	2.636363636

Figure 43 Mean Clicks Each Websites Open Domain Use case Evaluation[119]

The highest difference in the mean value was in the 'Killarney' website. To identify if the difference in the mean values are statistically significant a t-test assuming equal variances with a hypothesised mean difference of 0 was conducted. The corresponding p-value resulted in 0.169 for 'Health and Fitness', 0.003 for 'Killarney' and 0.448 for Dingle concluding that there is a significant difference between the mean click values of 'Killarney' in both cohorts. A reason for this difference may lie in the fact that each website presented a different amount of links the participant could choose. On the 'Health and Fitness' website all five links were relevant. The Killarney website participants were presented with 14 links of which 4 were relevant. In the Dingle website 8 links were displayed of which 2 were relevant. The significant difference in the decrease of clicks in the cross-over from the first to the second website can be judged as evidence as a positive impact of CSP on the users performance.

Finally, cumulative data was analysed. Cumulative observations were not directly comparable with the observations above due to using all three website datasets as one

[119] For illustration reasons a 90% confidence value was chosen. Values related to the 95% confidence interval are provided in Table 9 on page 302 in the appendix.

dataset. Therefore the following comparison only serves investigative purposes. Overall the average values resulted in 10.42 clicks (median 9) for the CSP enabled websites and 14.3 clicks (median 12) in the CSP disabled websites (Figure 44).

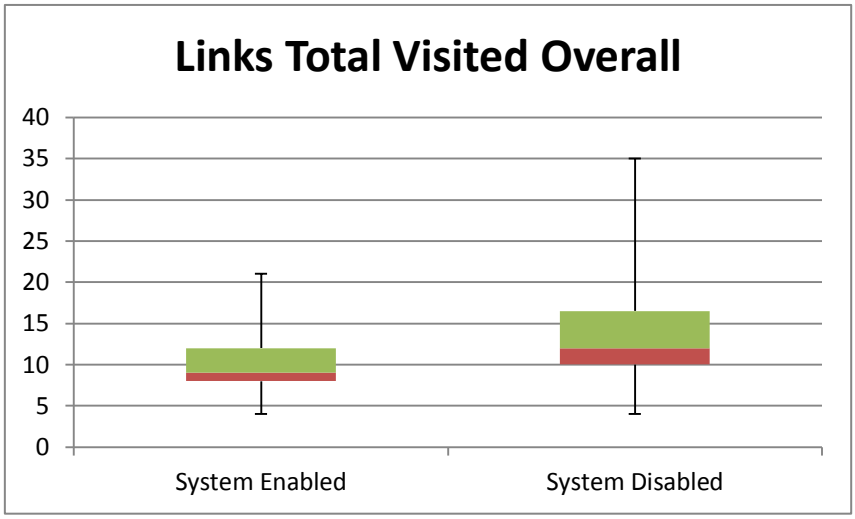


Figure 44 Mean Overall Clicks Open Domain Use case Evaluation

This result support the positive results above, which is also reflected in studying Killarney and Dingle without the initial (model training) website (Figure 45) with an average of 5.33 (median 5) clicks vs. 8.43 (median 7) clicks.

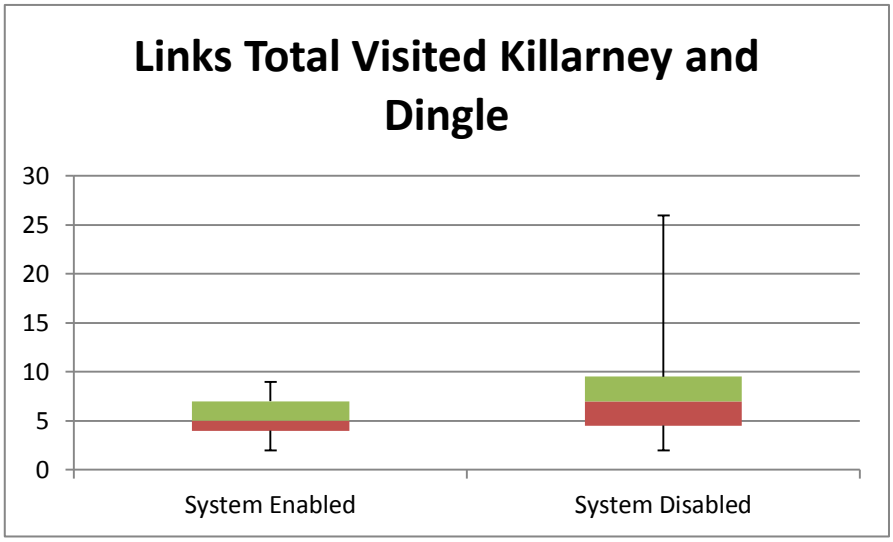


Figure 45 Mean Clicks Killarney and Dingle Open Domain Use case Evaluation

Task Completion Time

The overall task completion time (in seconds) within the CSP enabled cohort was 23100 and within the CSP disabled cohort 21960. The overall averages resulted in 1004 seconds vs. 954 seconds. To gain a more in-depth insight into the task completion time the overall mean times were compared by using a line chart (Figure 46).

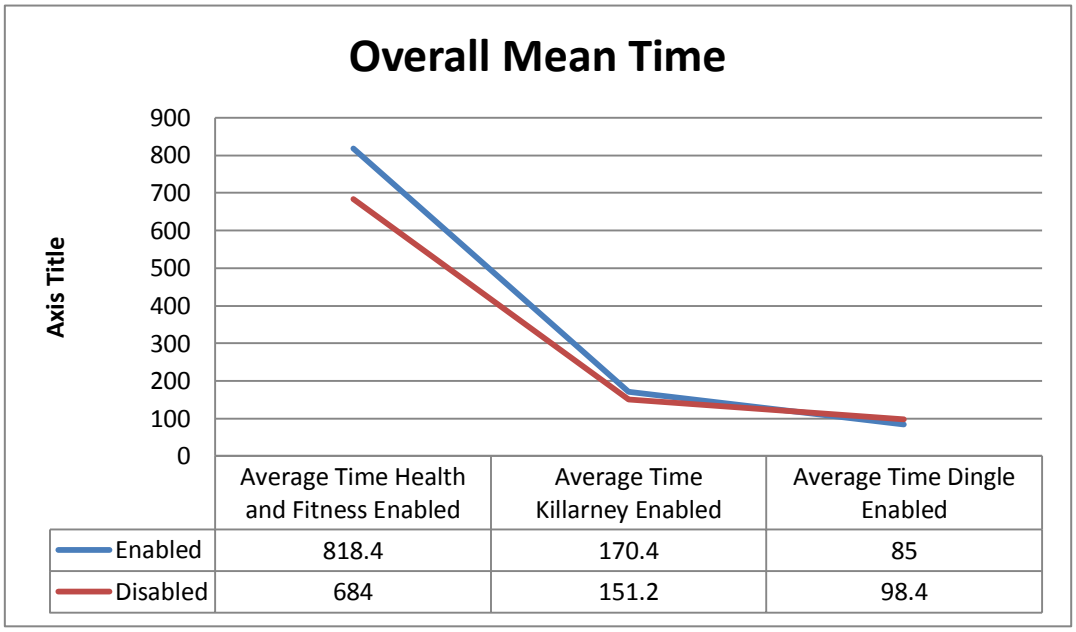


Figure 46 Overall Mean Times Open Domain Use case Evaluation

Interestingly, the gap between the means decreases. This trend is important and possibly indicates a performance increase of the introduced CSP approach as the users continue browsing across separate websites. Most encouraging is the decrease within the last website. To gain a better understanding, Figure 48 focuses on the last two websites (without the first – model training – website).

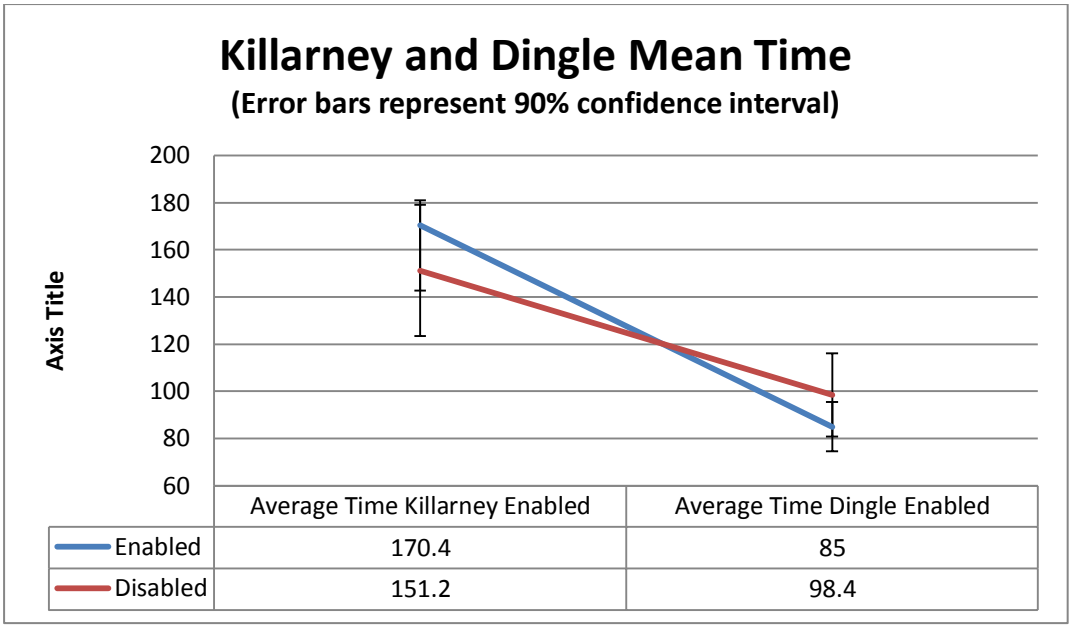


Figure 47 Mean Time Killarney and Dingle Open Domain Use case Evaluation

This decreasing trend is also visible in following box plot (Figure 48) indicating the difference in value distribution within the percentiles and the differences in minimum and maximum values.

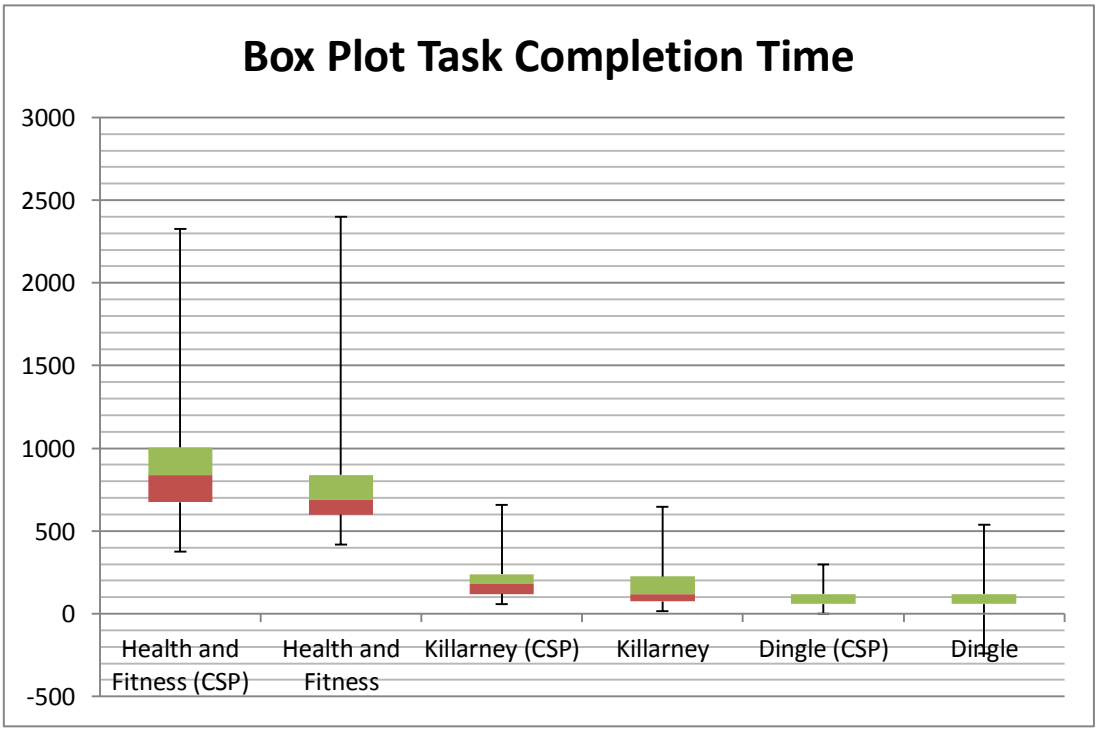


Figure 48 Box Plot Task Completion Time Open Domain Use case Evaluation

5.5.5. Results Self-Reported Feedback

To gather the self-reported feedback, the participants were asked to provide answers to 26 questions. The questionnaire had the objective of assessing the perceived appropriateness of the CSP approach, and whether it assisted users in addressing information needs in open domains. For this the participants were asked direct and indirect questions. Direct questions addressed key aspects of the CSP approach (e.g. perceived difficulty in browsing) and indirect questions assessed cognitive indicators, such as the perceived motivation and frustration.

The questionnaire was presented on a custom made online interface. The participants were presented with a slider to provide the answer on a scale from 0-100. Each participant was presented with visual markers at 0 for strongly disagree, 25 for disagree, 50 for neutral, 75 for agree and 100 for strongly agree. Yes and no questions were binary (0 for yes and 100 for no).

To assess the overall perceived mood of the participants both the perceived effect of the CSP approach on frustration and motivation was queried (Figure 49 and Figure 50).

The result indicated no statistically significant difference between both cohorts although with the average being slightly higher for the participants assessing the CSP enabled websites (averages 15.28 vs. 19.46 and median 15 vs. 11). Interestingly the max value (without outlier) reaches the maximum of 100 in the results related to the CSP disabled websites. The low level of frustration coincided with a high-level of motivation with the average of the CSP enabled websites feedback slightly higher (averages 80.29 vs. 74.93 and median 85 vs. 76).

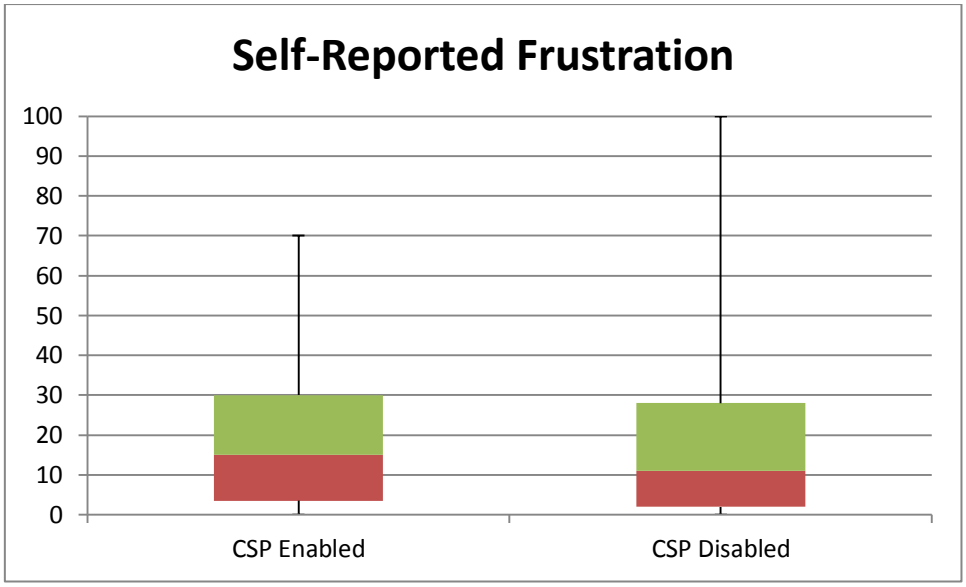


Figure 49 Self-Reported Frustration Open Domain Use case Evaluation

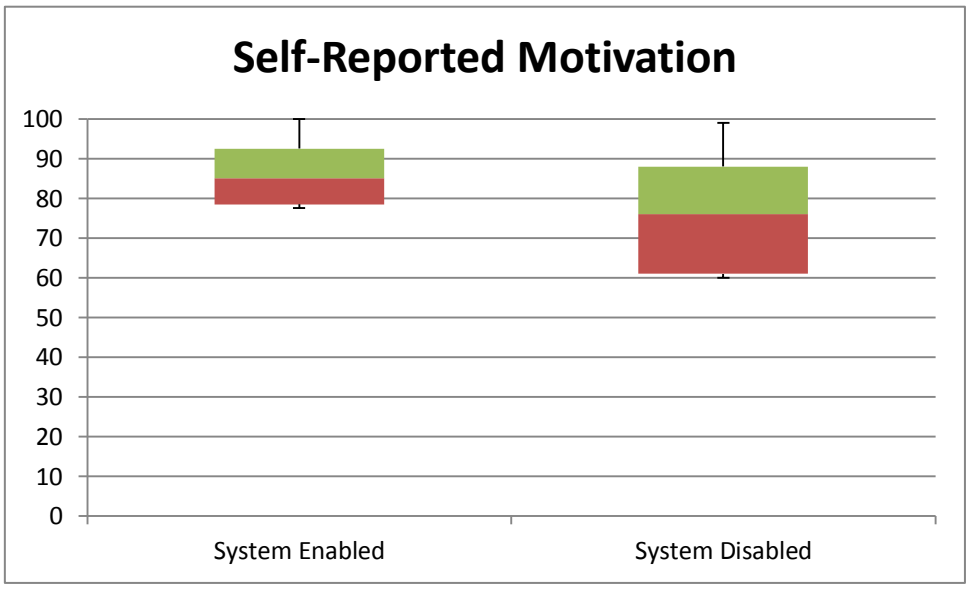


Figure 50 Self-Reported Motivation Open Domain Use case Evaluation

To assess if the task was appropriate the participants were asked if the task reflected a real-world cross-site browsing task. Most users agreed with similar means in both cohorts (averages 75.52 vs. 76.65 and median 76 vs. 77). This result is encouraging and indicates that the participants were not negatively biased due to the pre-set nature of the task and the unnatural browsing environment in a lab setting (Figure 51).

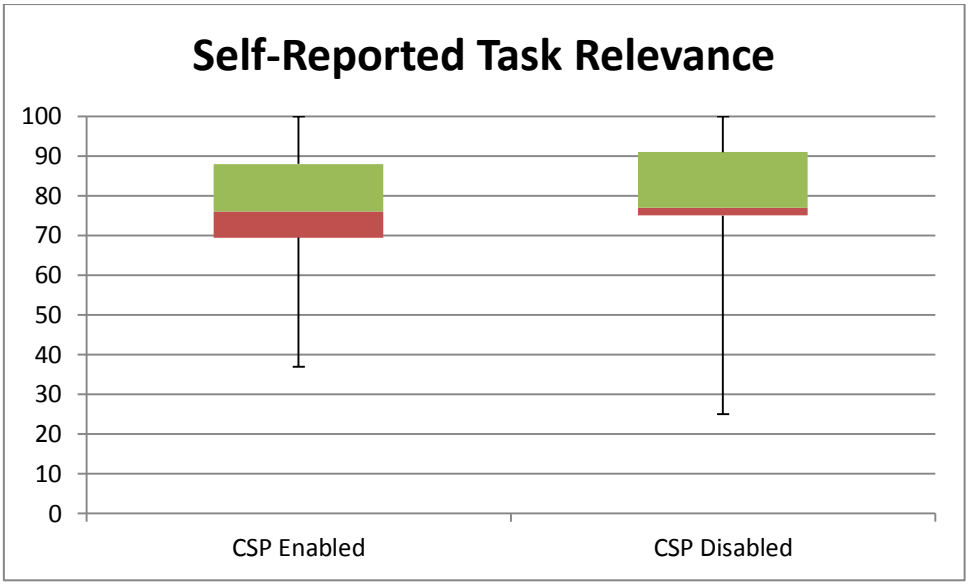


Figure 51 Self-Reported Task Relevance Open Domain Use case Evaluation

To assess the appropriateness of the CSP approach participants were asked to rate the perceived influence of system guidance (Figure 52). As mentioned in the experiment setup, no participant was informed about the CSP approach prior to the experiment. The result is therefore encouraging (average 79.55 vs. 59.41 and median 78 vs. 58) with the mean being **significantly** higher than the CSP disabled (p-values of 0.0006)[120]. This result can be interpreted, as a key finding in relation to the participant's perceived appropriateness of the approach to assist in cross-site information needs.

[120] To identify if the difference in the mean values are statistically significant a t-test assuming equal variances with a hypothesised mean difference of 0 was conducted.

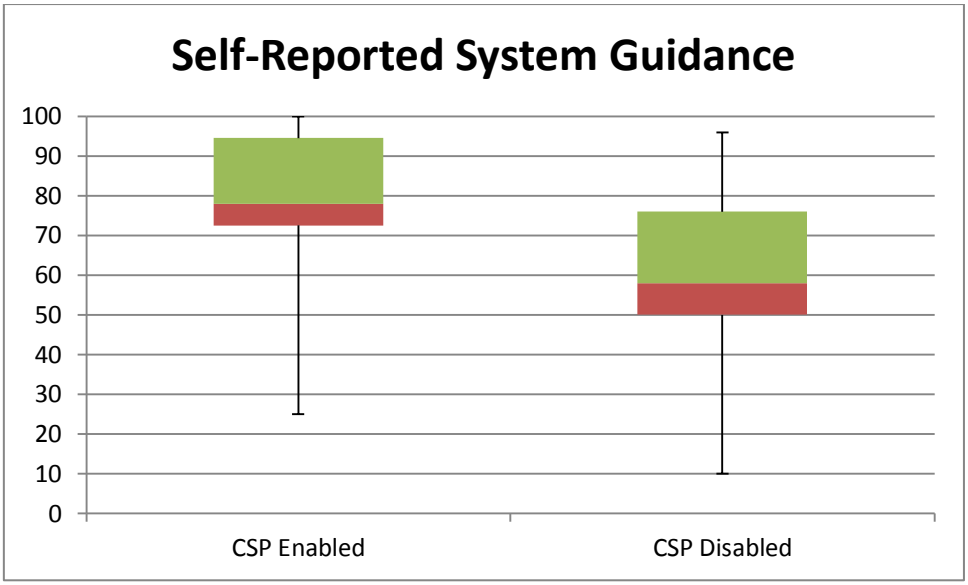


Figure 52 Self-Reported Usefulness of System Guidance Open Domain Use case Evaluation

A further encouraging result was found in the question related to the difficulty in browsing (Figure 53). The answers result in a **significant** difference towards the CSP enabled system setting (p-value 0.0216) with an average of 24.07 (media 25) in the CSP enabled based feedback and 36.52 in the CSP disabled 36.52 (media 38).

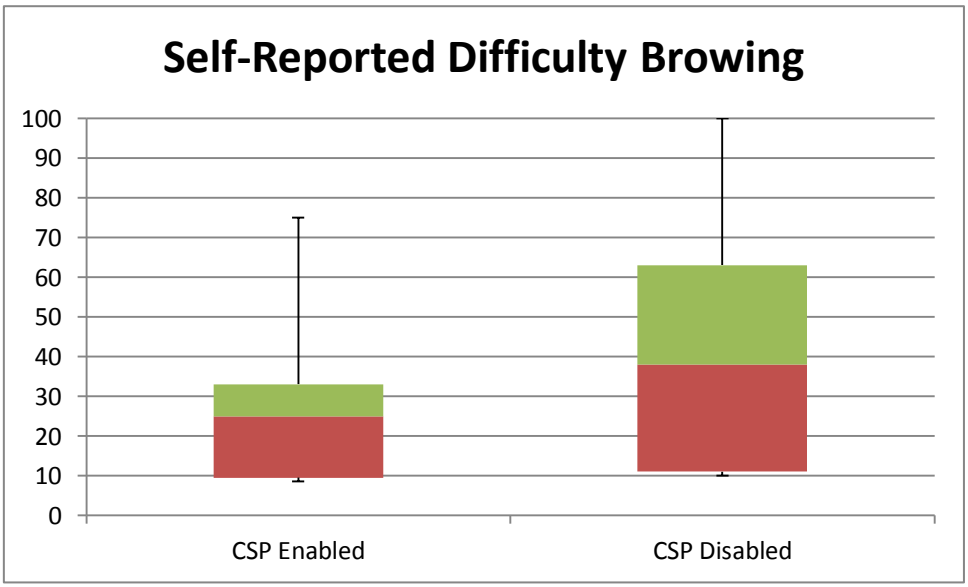


Figure 53 Self-Reported Difficulty in Browsing Open Domain Use case Evaluation

In relation to the perceived performance both cohorts indicated a high average (79.96 vs. 74.65) and mean (73.5 vs. 70) although without statistical significant difference (Figure 54). Even though this result does not indicate a distinct performance improvement towards the

CSP approach it does indicate that most participants were self-confident in performing the task.

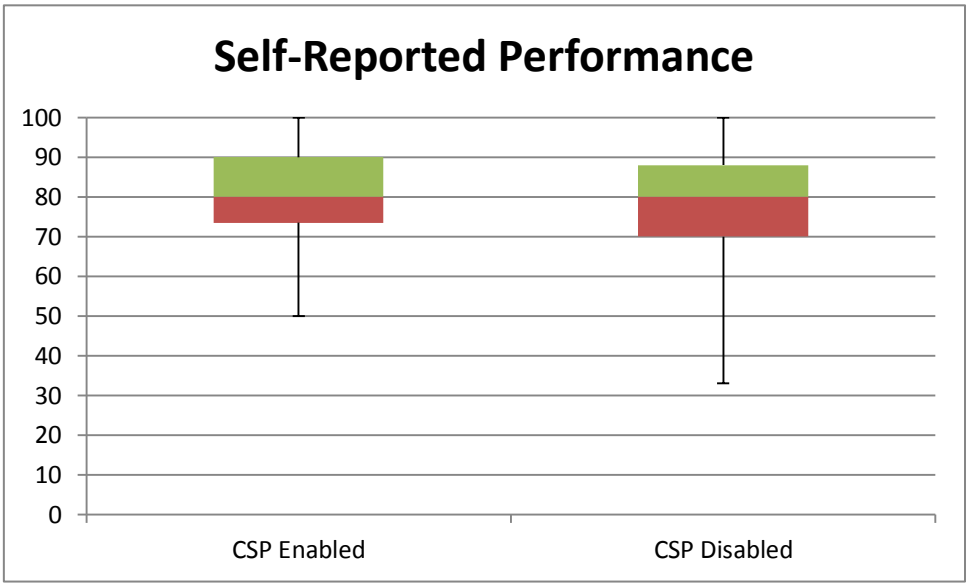


Figure 54 Self-Reported Performance Open Domain Use case Evaluation

Finally, an interesting result was provided by the participants in the question related to privacy concerns (Figure 55). Here, the cohort assessing the CSP enabled websites indicated a higher average concern than the cohort assessing the CSP disabled websites (59 vs. 55.79 and mean 46 vs. 48). This is surprising given the fact that this question does not directly relate to CSP or the usage of personalisation in general. In relation to this research the result can be interpreted as a higher awareness for the CSP approach within the cohort assessing the CSP enabled websites. Furthermore some participants indicated in the open interview at the end of the experiment that they felt watched and/or guided – a perception that might make users more curious about the source or motivation of the approach. Some participants stated they started to trust the guidance after a while and that they look for the visual indication first when accessing a website.

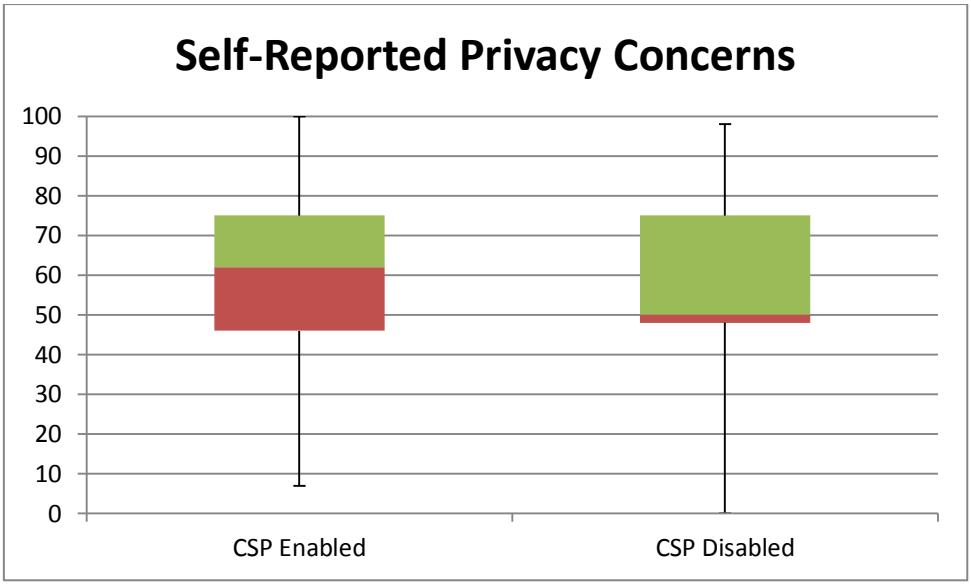


Figure 55 Self-Reported Privacy Concerns Open Domain Use case Evaluation

Overall, it can be concluded that the self-reported results are encouraging especially considering that some results indicate statistically significant differences in the mean values towards the introduced CSP approach.

5.5.6. Further Findings

Finally, a brief discussion is presented in relation to additional findings, including a brief summary of the different categories of participants and a discussion about task success.

The analysis of task completion time in different categories resulted in mostly overlapping results with a slight bias towards an increased time in most categories in the CSP enabled websites (Figure 56).

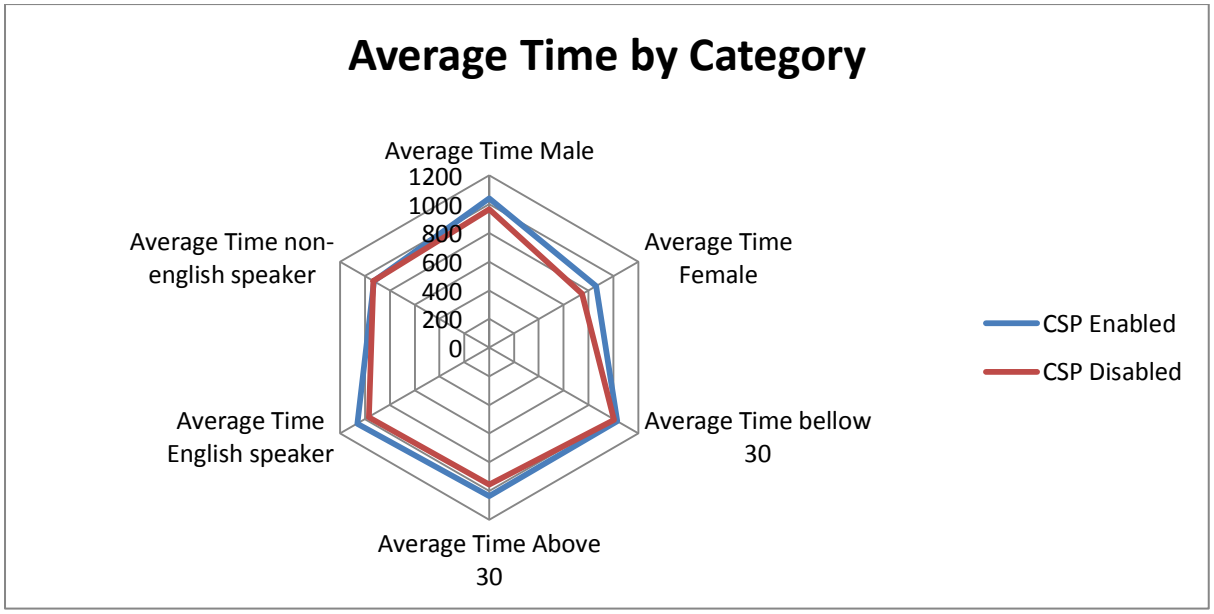


Figure 56 Overall Average Time by Category Open Domain Use case Evaluation

Analysing the second and third website (Killarney and Dingle) indicated lesser task completion time in the CSP enabled websites (Figure 57). It can therefore be concluded that the age, gender or language of the participants didn't have a significant effect on the findings.

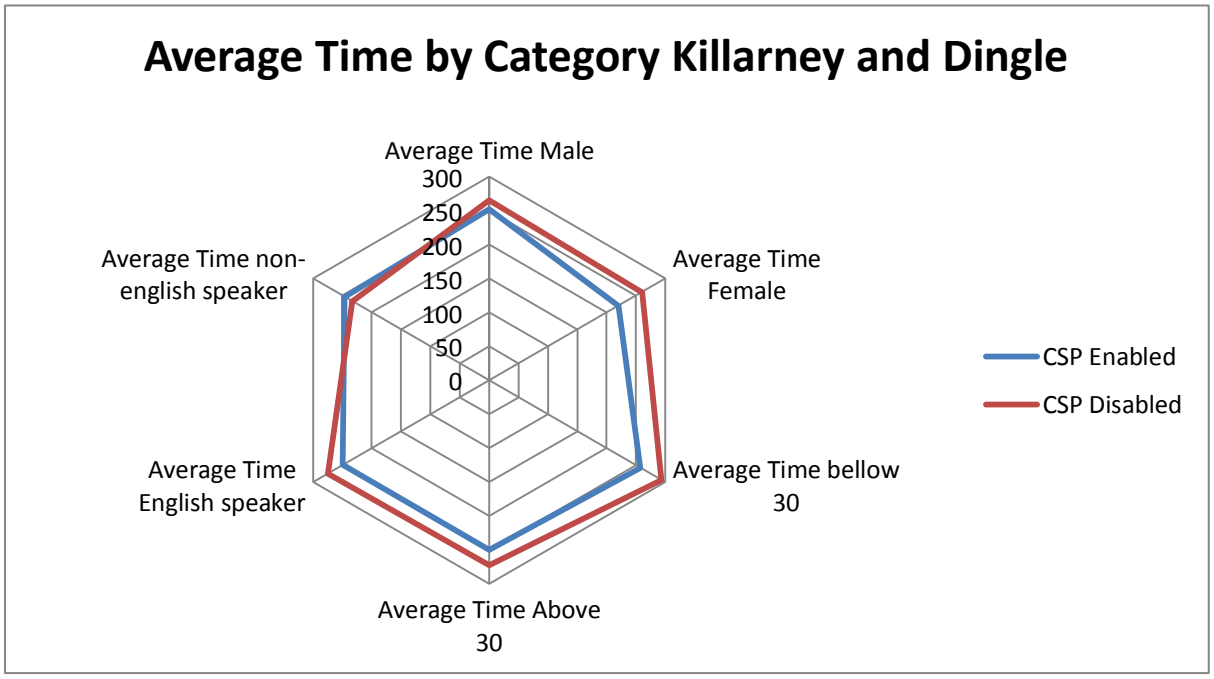


Figure 57 Average Time per Category Killarney and Dingle Open Domain Use case Evaluation

An analysis of the task completion rate and task success was conducted. In relation to this it can be concluded that the task completion rate was above 95%, which is possibly based on this experiment being lab based. In relation to task success the answers provided by the participants were analysed. Here, similar to the completion rate, most participants had a high-level of success (a fact that again can be argued is due to the lab based experiment environment). Furthermore the word count[121] indicated no evidence that one cohort would have required more time due to an increase in the time required to provide the task answers.

5.5.7. Discussion

The three objectives of this second experiment as introduced in section 5.5.1 were

1. To investigate the applicability of Cross-Site Personalisation in **open domains**.
2. To investigate the system requirements to consider the user's **privacy concerns**.
3. To investigate the engineering effort on the underlying website to ensure minimal integration effort of Cross-Site Personalisation into websites.

The first objective was assessed through user performance and self-reported metrics. The user performance indicated a trend in the decrease of both time and clicks on the third website browsed by the user. Moreover, it can be concluded that this is an indication for performance improvement the longer the participants used the system.

In relation to the user's self-reported feedback the difference was much clearer, with significant differences in the user's perceived guidance and perceived difficulty in browsing. Both answers, together with results related to motivation and frustration, resulted in a perceived relevance of the introduced CSP approach. Furthermore based on the possibility to conduct an unstructured interview it was possible to receive more insight in the usefulness of the CSP approach. One interesting aspect was that participants felt guided by the link annotations and after a while would start trusting it. In relation to this participants stated that they would first look for the link indication when coming to a new website. Even

[121] In the CSP enabled system setting, 6 participants had a low (<50 words) word count, 15 a middle word count (50-100 words) and 3 a high word count (>100 words). In the disabled setting 8 participants have a low word count, 11 a middle word count and 4 a high word count.

though not directly assessed this feedback provides a valuable indication that CSP may reduce the users perceive information overload when arriving to a new website.

The second objective in considering the users privacy concerns was technically addressed by ensuring the different websites do not receive any information about the browsing behaviour of the users on other websites. It should be noted that the users were not made aware of this feature it therefore remains a technical functionality.

Related to the third goal, the usage of WCMS module extensions ensured that website publisher only has to deploy a plug-in extension. Furthermore the approach taken by providing a third party service to interface with the websites on behalf of the user has proven to be both accepted by the participants, and technically feasible. All components within the approach, including the website modules and the interaction between the service and the websites, performed well and without any difficulties.

Finally, the three shortcomings identified in the first experiment and addressed by this second experiment are discussed.

The first shortcoming of using link-based visual assistance that were too subtle was addressed through a user study, and by ensuring that the visual assistance is consistently at a set position (within the menu structure) in all CSP enabled websites.

The second shortcoming of users, privacy concerns, was addressed by not sending the terms within the informed decision, but by adding link relevancies to a set list of links. This ensures that interfacing websites receive no information related to the user's cross-site information needs.

Finally, the third shortcoming related to missing open feedback from evaluation participants was addressed by conducting the entire experiment in a laboratory setting. Related to this it can be stated that the applied evaluation setup doesn't overcome all weaknesses in user focused web studies. It can be argued that the pre-setting of a browsing task is not a natural representation of the user's information needs in open domains. Overcoming this shortcoming however would require an open web deployment of the approach.

Overall, the outcome of second experiment can be stated as success. This observation is based on the positive indication of both the appropriateness and the perceived usefulness of the introduced approach. Furthermore the findings of the evaluation encourage future research and experiments related to Cross-Site Personalisation.

5.6. System Related Evaluation Considerations

Even though no direct system evaluation was conducted, in this section a focused overview of system specific considerations is discussed, with the goal to validate the technical appropriateness of the introduced approach. Specifically three areas are discussed:

(1) Technical evidence that the approach functioned successfully by discussing how the components within the overall approach fulfilled the design requirements;
(2) a discussion about system performance on both the CSP Service and the interfacing websites and finally,
(3) a brief discussion about the interconnection of both use case implementations, and how they addressed specific technical challenges related to CSP.

In relation to the first area the introduced approach was designed as a third-party service interfacing with websites through a RESTful API. To enable simple integration of CSP into websites module extensions were developed and tested. The role of the extensions was to facilitate the communication with the CSP Service and to provide non-intrusive navigational assistance based on an informed decision from the CSP Service. To test the technical appropriateness of the approach, two prototype systems were developed and evaluated. Both systems relied on the same high-level design introduced in the design chapter of this thesis. The main responsibility of the CSP Service was to generate an informed decision based on the user's cross-site information needs, which allowed interfacing websites to provide CSP to the user. The second important technical component of this research was the development of Web-based Content Management System (WCMS) module extensions. The role of the extensions is to interface with the CSP Service's API (sending user information based on browsing activity and receiving an informed decision) in order to provide visualise navigational guidance based on link annotations. The design and implementation of the overall approach was based on two use case based prototype implementations. Both implementations were separate and addressed two separate use cases (in closed and one in open domains), but by addressing the same set of high-level design requirements and by implementing the same high-level technical architecture (described in chapter three of this thesis), they were interconnected.

To ensure generality of the approach, the usage of different technology stacks was investigated and applied. For the first use case investigating CSP closed domains a Java based infrastructure was applied. The main rational was that closed domains usually consist of enterprise level websites interfacing with Java based backend systems. Therefore the CSP Service of the first use case based prototype was developed and deployed in Java. The second use case, investigating CSP in open domains, followed the current web paradigm of LAMP (Linux/Apache//MySQL/PHP), resulting in the CSP Service being implemented based on PHP. Both implementations performed without any major problems.

Based on the evaluation outcomes it can be concluded that in both cases the generation of an informed decision, based on the user's cross-site information needs, and the visualisation of the informed decision, was technically appropriate and successfully applied. This is based on the both the user's performance and the user's self-reported feedback being positive in both use case implementations.

In relation to the second area related to the performance implications of the approach, both prototype implementations were designed and implemented to ensure that the performance (page load time) of the interfacing websites is not affected, and to ensure that interfacing websites are not affected if the CSP Service fails to communicate with the websites. For both cases, coding procedures were put in place to ensure that the website can continue to function in the case either the CSP Service does not respond or the websites loading time is affected. Based on the non-intrusive notion of the introduced visual assistance, the user is still able to access the website if the CSP Service fails. A further performance based consideration was placed on the information exchanged between the websites and the CSP Service. Here, emphasis was placed on exchanging as little information as possible.

Finally, in relation to the interconnectedness of the use cases, both prototypes were developed in sequence and based on an iterative design process. The shortcomings identified in the first prototype development were consequently addressed in the second prototype implementation. Therefore the second prototype implementation is introduced as significantly more technically sophisticated. The reasons for this lies not only in the interactive nature of the design, but also in the fact that the second use case investigating

CSP in open domains is significantly more complex. This complexity is based on the highly dynamic nature of the open web. This is reflected by implementing the second CSP prototype as flexible infrastructure allowing the simple adding and removing of websites to and from the CSP browsing space.

It can therefore be noted that the technical appropriateness of the approach has been successfully validated by addressing the high-level design requirements (directly derived from the State of the Art review) and by applying continuous and rapid development cycles that reflected user-feedback and evaluation results.

5.7. Concluding Discussion

The evaluation chapter of this thesis discussed the findings of the different case study based evaluations, resulting in two different but interconnected prototype implementations for CSP. Each section concluded with a discussion and therefore much is not repeated in this concluding discussion. This section focuses on re-assessing in how far the high-level design affordances introduced in chapter three of this thesis were addressed and successfully evaluated by the two prototype implementations. Furthermore this concluding discussion seeks to address the Research Question, objective, goals and research challenges of the thesis as stated in the introductory chapter.

As mentioned above, this thesis introduces two interconnected prototype implementations. Both prototypes were assessed in two separate, but interconnected, experiments to evaluate the approach to assisting users in addressing information needs that span individually hosted websites. Furthermore a two-phased user study was conducted and discussed (5.4).

The first experiment assessed the appropriateness of CSP in closed domains (5.3). The use case was based on industry needs to provide CSP in customer care in which users browse corporate (structure) and user generated (unstructured) content. The content was supplied by Symantec and related to the Norton 360 product line. The second experiment focused on the appropriateness of CSP in open domains (5.5). It was based on tourism and used live content supplied from Board Failte Ireland.[122] To ensure real-world applicability, both use cases were designed to ensure all participants could relate to the tasks. The second prototype implementation, even though focused on an open domain, was designed to be applicable to both open and closed domains.

The results presented in this chapter provide evidence that the introduced approach assists users in addressing information needs that span independently hosted websites and that the design and implementation decision are appropriate, both from a user perspective and from a technical design perspective. Therefore it can be concluded that the discussion of the

[122] http://www.failteireland.ie/

evaluation outcomes within this evaluation chapter directly validates and confirm the Research Question of this thesis.

The following table (Table 7) discusses the appropriateness of the design and implementation decisions made to address the high-level requirements derived from the State of the Art (Table 2 on page 73). Furthermore in relation to the implementation details, the following table seeks to discuss the outcome of both use cases:

Section	**Concept Description**	**Component** (Figure 5 page 68)	**Evaluation Outcome**
3.1.1	A CSP approach should propose assistance that ensures the user is not hindered in their current browsing.	WCMS Module Extensions	Users indicated the chosen visual assistance as not hindering their browsing in both use case evaluations. (Figure 22 and Figure 42)
3.1.1	A CSP approach should have a unified understanding of the user's browsing space by creating a shared conceptualisation.	Term Identification Component and User Profile Component	Appropriateness was not directly evaluated. The design and implementation decision followed common practise. Users indicated no impact on performance in both use cases. (Figure 34)
3.1.1	A CSP approach should apply a user modelling technique that does not	WCMS Module Extensions and Term	Appropriateness was not directly

	interrupt the user's browsing. It should therefore be implicit by not requiring the user to explicitly participate in the identification, collection and management of information needs.	Identification Component	evaluated. The design and implementation decision followed common practise. Users indicated no impact on performance in both use cases. (Figure 36 and Figure 42)
3.1.1	A CSP approach should honour the user's privacy needs by ensuring the different websites the user is browsing do not receive any information about the user cross-site information needs.	Strategy Component	Appropriateness was not directly evaluated. The design and implementation decision followed common practise. Users indicated no impact on performance in both use cases. (Figure 55)
3.1.2	A CSP approach should be able to utilise existing content models to create a shared conceptualisation across different websites.	Term Identification Component	Appropriateness was indirectly directly evaluated. Based on the overall positive

			impact (Figure 52 and Figure 54) of the approach in both use cases it can be concluded that the term identification process was appropriate to enable CSP in closed domains.
3.1.2	A CSP should be able to use additional services and tools for term identification if necessary.	Term Identification Component	See above
3.1.2	Depending on the domain (open or closed) the CSP approach should be able to identify terms of content that is added or updated to website as soon as possible.	WCMS Module extensions and Term Identification Module	See above
3.1.3	A CSP approach should provide non-intrusive assistance techniques to ensure the user is not hindered in freely browsing.	WCMS Module Extensions	Users indicated the chosen visual assistance as not hindering their browsing in both use case evaluations (Figure 22 and Figure 42)
3.1.3	A CSP approach should affect the website as minimally as possible.	WCMS Module Extensions	Appropriateness was indirectly directly

			evaluated. Based on the overall positive impact (Figure 52 and Figure 54) of the approach in both use cases it can be concluded that the term identification process was appropriate to enable CSP in closed domains.
3.1.3	A CSP approach should not negatively influence core areas of the website (loading time etc.).	WCMS Module Extensions	See above
3.1.3	A CSP approach should be based on a design that allows simple integration and interfacing with existing websites.	WCMS Module Extensions	See above
3.1.3	A CSP approach should ensure that the communication between the CSP approach and the independent websites is flexible and does not depend on the websites technology stack.	Interface Layer	See above
3.1.3	A CSP approach should ensure device and browser independence.	Interface Layer	See above

Table 7 Design Requirements Addressed by the Evaluation

6. Conclusion

This thesis introduced a novel approach to assist users in addressing information needs that span independently hosted websites. This innovative approach aimed to addresses an identified gap in the State of the Art relating to the research fields of Adaptive Hypermedia and Web Personalisation. This gap refers to current approaches that only assist users within isolated websites and not across independent websites. This approach was introduced as Cross-Site Personalisation (CSP), facilitated through a third-party service and corresponding Web-based Content Management System (WCMS) extensions.

This chapter discusses the objectives and achievements of this thesis. Furthermore the contribution to the State of the Art is discussed. Finally, it concludes with a discussion of future work in relation to CSP.

6.1. Achievements

The research question of this thesis as stated in chapter one was:

Can users be assisted in addressing information needs that span independently hosted websites; and consequently, how appropriate might that assistance be?

Based on this research question, the following objective was defined for this thesis:

To address the above research question, the objective of this thesis is to introduce and evaluate a novel Cross-Site Personalisation (CSP) approach to assist users in addressing information needs that span independently hosted websites.

To address this objective three goals and related research challenges were stated:

1. *To **identify** the key requirements, techniques and impact of current Web Personalisation and Adaptive Hypermedia techniques in assisting users in addressing information needs that span independently hosted websites.*
2. *To **design and develop** a Cross-Site Personalisation approach to support users in addressing cross-site information needs.*
3. *To **evaluate** the appropriateness of the developed approach based on a series of case-study prototype implementations and experiments.*

The three stated research challenges were -

1. *To investigate the applicability of Cross-Site Personalisation in **open and closed domains.***
2. *To investigate the system requirements to consider the user's **privacy concerns**.*
3. *To investigate the **engineering effort** on the underlying website to ensure minimal integration effort of Cross-Site Personalisation into websites.*

In this section the research question, the three goals and related challenges are discussed.

Research Question

The overall research question – that asked: *How can users be assisted in addressing information needs that span independently hosted websites; and consequently, how appropriate might that assistance be?* –was addressed by introducing the design, implementation and evaluation of a novel approach defined as **Cross-Site Personalisation**. The overall approach consists of a third-party service providing an informed decision to websites about the user's cross-site information needs. Module extensions for Web-based Content Systems were developed with the goal to enable non-intrusive assistance navigational assistance. The approach (both service and module extensions) was evaluated through objective assessment, and based on scientific methodologies. The resulting user-performance and user-feedback data confirmed that the introduced approach can assist users in addressing cross-site information needs and that the assistance is appropriate.

Research Objective

The objective of this thesis - *to introduce and evaluate a Cross-Site Personalisation (CSP) approach to assist users in addressing information needs that span independently hosted websites* – was achieved through a sequence of use case studies and experiments. The experiments focused on two comparative and interconnected evaluations studying Cross-Site Personalisation in open and closed domains. Both use cases were based on real-world content and tasks. The outcome of both use cases prove that it is possible to assist users in addressing information needs across independently hosted websites. This conclusion is furthermore presented in several peer-reviewed publications, a full patent application,

three successful commercialisation grants and the on-going live web deployment and usage of the approach by a university spin out project called wripl[123]. Furthermore this thesis resulted in three related and completed MSc. projects.

Research Goals

*Goal 1: To **identify** the key requirements, techniques and impact of current Web Personalisation and Adaptive Hypermedia techniques in assisting users in addressing information needs that span independently hosted websites.*

To address this first goal, a State of the Art review was conducted. This review was divided into three sections: Web User Dimension, Web Content Dimension and Web Applications Dimension. The section related to the Web User Dimension identified and discussed information needs. The focus was placed on understanding how information needs originate and how such needs can be addressed. This was important in order to construct an understanding of cross-site information needs and how users seek to be supported in addressing such needs.

The web content section focused on semi-structured and unstructured content. Semi-structured content usually consists of a pre-set structure represented through a specific standard or description. Unstructured content does not have an underlying structure. The significance of this discussion in relation to the research of CSP is the difference in strategy related to identifying terms describing the content. Based on the inherent structure of semi-structured content, it is possible to apply semi-automatic term identification approaches (e.g. machine learning) for more accurate term identification. To extract terms for unstructured content Name-Entity-Relationship (NER) approaches can be applied. Both content types are not distinct. It is possible that websites host both types of content. In relation to this research this distinction was important to understand and motivated possible use cases in which Cross-Site Personalisation could be applied.

The final section of the State of the Art review (Web Application Dimension) discussed Web Personalisation from three different perspectives. The first related to applying Web

[123] http://www.wripl.com

Personalisation in a non-intrusive manner to ensure the user is not limited in freely browsing. The second, focused on non-invasive integration of Web Personalisation to websites, so that Web Personalisation techniques do not interfere with the website in a way the website publisher may not want. Finally, the third perspective emphasised interoperability in enabling Web Personalisation to be applied across separate websites.

The importance of this discussion was to develop an in-depth understanding of how an approach that assists a user in addressing information needs across different websites can balance both the needs and preferences of the user (non-intrusive assistance) and the needs and preferences of the website publisher (non-invasive integration) and how this balance (equilibrium of personalisation) can be applied/maintained across separately hosted website (interoperability). Based on the novelty of this approach, no comparative study of different CSP approaches was possible. However, approaches related to CSP were discussed independently allowing the identification of relevant aspects and challenges. The State of the Art review addresses this research goal and is documented in chapter two of this thesis.

Goal 2: *To **design and develop** a Cross-Site Personalisation approach to support users in addressing cross-site information needs.*

To address this second research goal, a set of design requirements were derived from the State of the Art review. Based on the design requirements a high-level and two use case based designs were presented in the design chapter. The designs are based on an iterative development approach. First, high-level requirements were identified through the State of the Art review described in chapter two of this thesis. The State of the Art review was extended with a study of Web-based Content Management Systems (WCMS) and an initial user questionnaire to assess the needs and requirements for CSP. The result of this initial review and assessment was a set of design requirements resulting in a high-level design of the CSP approach. The high-level design was introduced as an approach to CSP based on a centralised third-party service that interfaces with the backend of the different websites the user browses. The interface is based on a RESTful API to ensure an abstraction from the websites specific technology stack or subject domain. A further element of the high-level design, are custom built WCMS module extensions. These extensions were implemented to support non-invasive integration of CSP to a website and to allow a website to offer

navigational assistance to the user in a non-intrusive manner. The CSP Service together with the Website module extensions forms the overall CSP approach which introduces a **CSP Browsing Space**.

To ensure privacy-sensitive browsing, the design introduces the concept of an **informed decision** that allows websites to provide visual navigational assistance without exposing information related to the user's cross-site information needs.

To address the different requirements four components were introduced. All four components are part of the CSP Service. The first component (Term Identification Component) addressed the identification of terms describing web content (3.2.2.1). A second component addresses the storing of the terms related to the user's information needs together with the relevancy of the terms (3.2.2.2). The third component is responsible for the creation of an informed decision based on the user's cross-site browsing interest (3.2.2.3). To increase user trust the user can access the information the CSP Service has stored through the fourth component responsible for user profile scrutiny (3.2.2.4).

The initial high-level design was used in two distinct, but interconnected, use cases. The first use case illustrated CSP in closed domains. Based on the outcome of the first use case evaluation, a second use case addressed CSP in open domains. The outcomes of this iterative process was two distinct, but interconnected, approaches to CSP based on the design requirements and high-level design.

*Goal 3: To **evaluate** the appropriateness of the developed approach based on a series of case-study prototype implementations and experiments.*

The third research goal, was addressed by focusing on the two introduced use case implementations. The iterative design nature of the development resulted in two distinct design instances. Furthermore both use cases addressed different information needs. The first addressed information needs that are more specific (fact finding) and requires subject specific websites that. The second use case addressed information needs that are not well

defined (fact gathering)[124] and where it may be necessary to browse websites with multiple subject domains.

Both use cases based prototypes were evaluated in comprehensive user focused studies. The first evaluation was taken by users in isolation from the evaluator conducting the experiment. The tasks in this first evaluation were related to Online Customer Experience (specifically customer care) and involved the browsing of two separate websites. The experiment was conducted with 60 participants and as between-subject design. The results confirmed the appropriateness of the approach, based on user-performance and self-reported feedback. A further indication, derived from the results, was that the introduced visual assistance might have been too non-intrusive, with several users stating that they were not able to find any difference (even though an introduction video was provided). This was addressed by splitting the second experiment into different phases inspired by the layered evaluation approach introduced by Paramythis et al. (2010).

Prior to a second use case based prototype evaluation within open domains, two interview based user studies were conducted. Both user studies were conducted as structured and open interviews. The first study focused on the visual assistance of CSP across different websites. This study resulted in the identification of visual assistance details that allow a consistent browsing assistance across independent websites. The second study focused on the accuracy of the informed decision by the CSP Service. For this the participants were asked to rate the relevance of links based on their (pre-set) browsing interest. This user relevancy was compared with the informed decision of the CSP Service. The result of this comparison was positive with most calculations overlapping. Interestingly many users tended to rate the links as either 0 or 10 and not using any value in between.

After concluding the two-focused user studies, a comparative use case based evaluation was conducted. For this study 60 users were sourced. The task was exploratory and based on fact gathering. Compared with the first comparative evaluation (closed domain) this experiment was conducted in a laboratory setting. The participants were left alone during the task and during the answering of task related questionnaires. After concluding the task,

[124] For a comprehensive overview of web information tasks note (Kellar et al., 2006b).

the evaluator conducted an unstructured interview to identify further challenges. Furthermore the experiment was conducted as between-subject design (each participant only evaluated one system setting) to ensure the effect of CSP in relation to the baseline setting (CSP disabled) could be studied.

The analysis of the self-reported feedback found statistical difference in the feedback towards the perceived appropriateness of the approach in relation to system guidance (Figure 52) and difficulty of browsing (Figure 53). In relation to the user-performance metric, a decreasing trend was identified in the time and clicks (links accessed) as a user browsed across independent websites. The overall results (self-reported and user-performance) were encouraging with most users indicating the appropriateness for CSP to assist users in cross-site browsing.

In relation to the engineering effort, the CSP approach was designed to enable simple integration with existing websites through module extensions and by offering a RESTful API. In relation to system performance, no user indicated any performance issues. This feedback illustrates that the usage of an IR based approach for the calculation of the informed decision is appropriate to provide consistent assistance across independently hosted websites. In relation to scrutiny, users indicated a strong interest for deeper insight into the information the service collected.

Research challenges

The research challenge of investigating the applicability of CSP in open and closed domains was addressed by conducting the two comparative user based evaluation described above. Both evaluations illustrated the applicability of CSP in both use cases (open and closed web).

The second research challenge of investigating the system requirements to consider the user's privacy concerns was addressed by conducting an initial user focused survey and by continuously addressing and refining privacy issues based on participant feedback. The resulting mechanisms in relation to privacy concerns ensure that the interfacing websites only received informed decision from the CSP Service, but no information on the needs and preferences of the user. Furthermore based on an introduced scrutiny component the user can view the information the service has collected on behalf of the user.

The final and third research challenge focused on investigating the engineering effort on the underlying website and to ensure minimal cost and time effort in deploying the introduced cross-site support mechanism. This was achieved by introducing module extensions for WCMS based websites. The design was focused on non-invasive integration and therefore ensured that only that the website publisher can enable and control the integration of the extension. Furthermore designing a RESTful API allows any type of website/application to interface with the CSP Service.

6.2. Contribution to the State of the Art

A novel approach to **assist users in addressing information needs that span independently hosted websites** is the primary contribution to the State of the Art made by this thesis. This research addresses an existing gap in the State of the Art related to a lack of seamless assistance across independently hosted websites. Current techniques and approaches focus either on assisting users in stand-alone applications/websites, or by providing means to interchange content and models known as the open-corpus (OC) problem (Henze and Nejdl, 2001). Currently no approach has been introduced that provides consistency and interconnected assistance across independent websites. The introduced approach addresses this gap by introducing a third-party service and Web-based Content Management System extensions that allow users to receive consistent (based on the user's cross-site information needs) assistance across independently hosted websites by means of navigational guidance. This approach is introduced as **Cross-Site Personalisation** and forms the main contribution of this thesis. Further contributions are to ensure the Cross-Site Personalisation applied is (1) **non-intrusive** for the user, (2) **non-invasive** for the website and (3) by introducing the concept of **informed decision** honours the privacy needs of the user. Furthermore by introducing Web Personalisation through Web-based Content Management System extensions, interoperability in the applied personalisation techniques is introduced to the research area of Web Personalisation.

These contributions have resulted in the following publications:

Koidl K., Conlan C., Wei L. and Saxton A.M. (2011) Non-invasive Browser Based User Modeling Towards Semantically Enhanced Personlization of the Open Web. FINA-3A: Semantic Web and Systems, AINA 2011, Singapore.

Koidl K., Conlan O., Wade V. (2009) Non-Invasive Adaptation Service for Web-based Content Management Systems (AH2009), International Workshop on Dynamic and Adaptive Hypertext: Generic Frameworks, Approaches and Techniques, Torino, Italy [.

Koidl, K., Conlan, O. (2008) Engineering Information Systems towards facilitating Scrutable and Configurable Adaptation, Fifth International Conference on Adaptive Hypermedia and

Adaptive Web-based Systems (AH2008), Springer Lecture Notes in Computer Science 5149, Hannover, Germany, 28 July - 1 August 2008, pp405-409.

This work provided the foundation for three completed MSc projects. This research has been filed as full patent application (US and UK) and has resulted in three successful funding proposals.

6.3. Future Work

Several potential areas were identified where the research described in this thesis could be extended and advanced. The main areas are discussed below.

6.3.1. User Profile Management and Social Media

The CSP approach presented in this thesis allows the user to view terms (model scrutiny) relating to the user's cross-site information needs. Through comments, several participants expressed their interest in actively engaging with their user profile. For this, future work could focus on the following aspects:

1. Provide insight on where the information was collected and used.
2. Enable users to add and delete terms within the user profile.
3. Enable users to actively manage terms within different 'interest sections'. These sections could be created by the user and allow the grouping of related terms.

One of the shortcomings identified in the State of the Art is related to users either rapid switching information needs or addressing several different information needs at the same time (e.g. through browser tab usage). A possible solution to address this shortcoming is to analyse the higher-level topic of a website. For example if a website relates to the topic of News and Media the CSP service could introduce a new 'interest section' relating to that topic and add all terms to that section. If the topic related section already exists, then the service could assign the terms into the appropriate section. This also allows the CSP approach to use terms from the appropriate interest. This however poses a higher-level challenge. Some information needs may require the browsing of websites with different topics. For example planning a holiday may also include seemingly unrelated website topics, e.g. related to a hobby an interest in historical artefacts and wanting to connect this interest with a holiday location. Overcoming this challenge may involve a more active participation of the user in managing their own profile.

Finally, user profile sharing could be introduced as a social dimension to CSP. Allowing the sharing of user profiles would enable users to share common information needs, such as exploring information about an upcoming holiday, with friends and peers. In addition to

direct social sharing, public sharing could be introduced to allow users to make parts of their profiles public and available for others to use. This makes most sense in mature user profiles that have been created over a longer time period and across several websites. A third social related functionality is related to social media applications. By facilitating the social graph of the user the pre-population of the user's profile might be possible (Abel et al., 2010).

6.3.2. Term Identification

In this thesis two different approaches to term identification were introduced. The first relied on in-depth design-time analysis of provided content through machine learning techniques. The second relied on an external service to allow a more open and flexible term identification. Both approaches could be extended. The first approach could be extended to utilise semantic web related information, such as RDFa[125] and Microformats[126]. This would rely on the websites using a specific metadata format or standard. Furthermore, based on the closed domain addressed by the first use case, the application of CSP could trigger a more stringent use of metadata by website publishers within a closed domain.

In relation to the second use case applied (open domain) the current design allows for the extension to add additional term identification services. The service used in this use case was OpenCalais. This service was chosen based on its wide adoption and support. Further services could be introduced to the CSP Service depending on the use case. Additional use cases, for example based on social media content, could use services that are more effective in identifying meaning and/or sentiment of social media related content (Rizzo, 2011). Finally, developments in Name-Entity-Recognition tools, such as Apache OpenNLP[127] and Apache Stanbol[128], should be further investigated as they would allow a reduction of a third-party service dependence within the approach.

[125] http://www.w3.org/TR/xhtml-rdfa-primer/

[126] http://microformats.org/

[127] http://opennlp.apache.org/

[128] http://stanbol.apache.org/

6.3.3. Information Need Detection

The approach introduced for information need detection relies on a server side integration of websites with the CSP Service. The two main reasons for this, is to ensure CSP being both device and browser independent and to allow the CSP approach to enable deeper and more reliable integration into the website. The main limitation of this approach is that CSP is only provided within the CSP browsing space. This browsing space limitation has benefits in terms of trust and privacy, however if a user browses to a website that hasn't integrated the CSP approach, the informed decision might be unrelated or out of date. A possible approach to avoid this, is to offer the user a browser plugin or mobile browsing app that allows the monitoring of the browsing behaviour outside of the confines of the CSP browsing space. The result would be a hybrid approach in which personalisations would be offered only on websites that interface with the service (for security and data privacy reasons) but ensure that the informed decision is accurate based on the broader collection of browsing information.

6.3.4. Flexible Intrusiveness

A core element of this research was to ensure that the introduced approach would ensure non-intrusive assistance for the user and non-invasive integration for the website publisher. These design requirements were derived based on the State of the Art analysis and user studies. The commonality of both aspects (non-intrusive assistance and non-invasive integration) is to balance the needs and requirement of the user and the website publisher (state of equilibrium). This thesis argues that the means to establish this balance has been introduced by CSP. Extensions to the CSP approach could allow this balance to be adjusted to either side. For example the user may opt into more intrusive guidance that in the most extreme could result in a fully personalised website. This is possible due to the approach integrating directly with the website and therefore enabling deeper personalisations. The main question that should be addressed for this extended work is who controls the approach? This thesis argues that the control of intrusiveness should be based on the user's information needs. This however may conflict with the website publisher's economic need. An example of this conflict could be a user seeking full intrusiveness i.e. a fully guided and personal website, but not seeking any links that are of commercial benefit to the website

publisher (such as promoted links, advertisement etc.). To overcome this shortcoming an adjustable or negotiable element could be introduced to the service that indicates the level of intrusiveness a user seeks and the level of invasiveness a website publisher allows. The CSP Service could then negotiate/create a balance (equilibrium) based on the pre-set requirements of both sides.

A further intrusiveness related extension could be to allow users to choose the visual navigation assistance used for the informed decision. In the current approach, link annotation is used. However, enabling the user to choose a different visual navigation assistance may require the permission of the website publisher. To ensure consistency across the separately hosted websites, each website would have to ensure the selected visual assistance by the user is provided on each CSP enabled website. This however may not be in the interest of the website publisher due to look and feel requirements. To overcome this potential conflict, visual assistance preferences could be added to the above mentioned negotiation approach.

6.3.5. Mobile and Localisation

The service based approach ensures that CSP could be applied on websites rendered on a mobile device. This assists mobile use cases in which a user browses CSP enabled websites on a stationary desktop and then continues to browse the CSP enabled websites on a mobile device. This seamless **Cross-Device Personalisation** could be extended by integrating CSP deeper into the mobile device through app development. This would allow the CSP service to receive information about the user's location, and potentially allow the integration with other app information to enrich the user's cross-site profile and cross-site personalisation. This extension might also include using the user's cross-site profile to provide location aware recommendations or by providing different apps with an informed decision.

A further, potentially mobile related, future work is in the context of localisation. Specifically the application of CSP across different language sets. This would allow users to avail of CSP assistance on websites hosting content in different languages. An example of this would be in the tourism domain. Here a non-English speaking user may seek to browse websites hosting in a local language and related to a specific hobby. The user could use CSP to seek visual assistance on English content based websites, e.g. a tourist location, that match the

user's hobby. To enable **Cross-Lingual CSP** two approaches could be studied. The first would translate all source (e.g. French) language based websites into one target language (e.g. English) and extract terms only from the target language corpus. The other, opposite approach, would be to extract terms in the source languages (e.g. French) and then translate the terms into one target language (e.g. English). Obviously the target language in both cases would relate to the user's mother tongue and/or language preference. Future work in this space relies strongly on the investigation of semi-automated machine translation and multi-language term identification.

6.3.6. Web Narratives

One of the main concepts introduced by Cross-Site Personalisation is the notion of an *informed decision* that allows websites to provide navigational assistance without compromising the user's privacy.

Two strategies were presented to create an informed decision. The first strategy (introduced within closed domains) applies term matching to determine the relevancy of links (3.3.3.3). The second, more sophisticated, strategy (introduced within open domains) applies Information Retrieval techniques to identify the relevancy of links (3.4.3.3).

As this thesis has proven, both approaches are appropriate in assisting the user in exploration and fact gathering based browsing. However, in tasks that require a more structured guidance and which may underlie a specific goal or objective, such as learning a specific topic, the approach can be extended towards a more narrative driven decision process. The notion of introducing a narrative has been successfully applied in Adaptive Hypermedia Systems (AHS) traditionally focused on the area of e-Learning (Conlan and Wade, 2004). Within AHS the usage of a narrative allows the sequencing of course material based on the course description of a pedagogical expert. This narrative can then be used to populate a course with content based on pre-defined concepts. By overlaying the content concepts with user model concepts a balance between user preferences (e.g. learning goals, current knowledge etc.) and narrative preferences (a defined sequence of the course) can be reached (Abbing and Koidl, 2006).

This narrative approach, as described above, could be applied within Cross-Site Personalisation. For this the Term Identification Component would require an extension toward extracting higher-level conceptual terms. Based on the learning example above, this can include the identification of terms such as '*Orientation'*, '*Exploration'* and 'Exam'. One other example is by difficulty level of the content, such as '*Beginner'*, '*Expert'* and '*Novice'*. A third example can be the question type addressed by the content: '*How to?*' or '*What is?*'.

The identification process of high level terms can be challenging and may require users to participant in a semi-automatic approach (Sah and Wade, 2010). Due to Cross-Site Personalisation introduced as a service approach, it could be envisaged that this identification process could be crowed-sourced or shared across the different users registered with the service.

The introduction of higher level terms to Cross-Site Personalisation would enable users and/or subject experts to construct Web Narratives that would enable a goal specific navigational guidance of the user across the web.

This chapter presented the achievements, contribution to the State of the Art and future work related to the research presented in this thesis.

Bibliography

Abbing, J., Koidl, K., 2006. Template approach for adaptive learning strategies, in: Proceedings of Adaptive Hypermedia. Presented at the AH 06, Dublin, Ireland.

Abel, F., Henze, N., Herder, E., Krause, D., 2010. Interweaving Public User Profiles on the Web, in: UMAP. pp. 16–27.

Abel, F., Herder, E., Houben, G.-J., Henze, N., Krause, D., 2013. Cross-system user modeling and personalization on the social web. User Modeling and User-Adapted Interaction 23, 169–209.

Abiteboul, S., 1997. Querying semi-structured data. Database Theory—ICDT'97 1–18.

Acerbis, R., Bongio, A., Brambilla, M., Butti, S., 2007. Webratio 5: An eclipse-based case tool for engineering web applications. Web Engineering 501–505.

Adkinson, W.F., Eisenbach, J.A., Lenard, T.M., 2002. Privacy Online: A Report on the Information Practices and Policies of Commercial Web Sites. Progress and Freedom Foundation Washington DC.

Adomavicius, G., Tuzhilin, A., 2005. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering 17, 734– 749.

Ahn, J., Brusilovsky, P., He, D., Grady, J., Li, Q., 2008. Personalized web exploration with task models, in: Proceedings of the 17th International Conference on World Wide Web, WWW '08. ACM, New York, NY, USA, pp. 1–10.

Ahuja, J.S., Webster, J., 2001. Perceived disorientation: an examination of a new measure to assess web design effectiveness. Interacting with Computers 14, 15–29.

Anand, S.S., Mobasher, B., 2005. Intelligent Techniques for Web Personalization, in: Mobasher, B., Anand, S.S. (Eds.), Intelligent Techniques for Web Personalization. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–36.

Arasu, A., Garcia-Molina, H., 2003. Extracting structured data from web pages, in: Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data. ACM, pp. 337–348.

Avison, D., Fitzgerald, G., 2003. Information systems development: methodologies, techniques and tools. McGraw Hill.

Bailey, C., Hall, W., Millard, D.E., Weal, M.J., 2002. Towards Open Adaptive Hypermedia, in: AH. pp. 36–46.

Bailey, C., Hall, W., Millard, D.E., Weal, M.J., 2007. Adaptive hypermedia through contextualized open hypermedia structures. ACM Transactions on Information Systems (TOIS) 25, 16.

Bates, M.J., 2002. Toward an integrated model of information seeking and searching. The New Review of Information Behaviour Research 3, 1–15.

Belkin, N.J., 1980. Anomalous states of knowledge as a basis for information retrieval. Canadian journal of information science 5, 133–143.

Belkin, N.J., Croft, W.B., 1992. Information filtering and information retrieval: two sides of the same coin? Commun. ACM 35, 29–38.

Belkin, N.J., Oddy, R.N., Brooks, H.M., 1982. Ask for information retrieval: part I. Background and Theory. Journal of Documentation 38, 61–71.

Berendt, B., Spiliopoulou, M., 2000. Analysis of navigation behaviour in web sites integrating multiple information systems. The VLDB Journal 9, 56.

Berghel, H., 1997. Cyberspace 2000: dealing with information overload. Commun. ACM 40, 19–24.

Berners-Lee, T., 2010. Long Live The Web. Scientific American 303, 80–85.

Berners-Lee, T., Hall, W., Hendler, J.A., O'Hara, K., Shadbolt, N., Weitzner, D.J., 2006. A framework for web science. Found. Trends Web Sci. 1, 1–130.

Bernstein, M.S., Ackerman, M.S., Chi, E.H., Miller, R.C., 2011. The trouble with social computing systems research, in: Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems. pp. 389–398.

Bizer, C., Heath, T., Berners-Lee, T., 2009. Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems 5, 1–22.

Bohnert, F., Zukerman, I., 2009. Non-intrusive personalisation of the museum experience. User Modeling, Adaptation, and Personalization 197–209.

Bra, P.D., Aerts, A., Berden, B., Lange, B. de, Rousseau, B., Santic, T., Smits, D., Stash, N., 2003. AHA! The adaptive hypermedia architecture, in: Proceedings of the Fourteenth ACM Conference on Hypertext and Hypermedia. ACM, Nottingham, UK, pp. 81–84.

Bra, P.D., Aerts, A., Smits, D., Stash, N., 2006a. AHA! Meets AHAM, in: Bra, P.D., Brusilovsky, P., Conejo, R. (Eds.), Adaptive Hypermedia and Adaptive Web-Based Systems, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 388–391.

Bra, P.D., Smits, D., Stash, N., 2006b. Creating and Delivering Adaptive Courses with AHA!, in: EC-TEL. pp. 21–33.

Brin, S., Page, L., 1998. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 30, 107–117.

Broder, A., 2002. A taxonomy of web search. SIGIR Forum 36, 3–10.

Brooke, J., 1996. SUS-A quick and dirty usability scale. Usability evaluation in industry 189, 194.

Brusilovsky, P., 1996. Methods and techniques of adaptive hypermedia. User Modeling and User-Adapted Interaction 6, 87–129.

Brusilovsky, P., 2001. Adaptive Hypermedia. User Modeling and User-Adapted Interaction 11, 87–110.

Brusilovsky, P., 2004. KnowledgeTree: a distributed architecture for adaptive e-learning, in: Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers \& Posters. ACM, New York, NY, USA, pp. 104–113.

Brusilovsky, P., 2008. Adaptive Navigation Support for Open Corpus Hypermedia Systems, in: Proceedings of the 5th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems. Springer-Verlag, Hannover, Germany, pp. 6–8.

Brusilovsky, P., Henze, N., 2007. Open Corpus Adaptive Educational Hypermedia. The Adaptive Web 4321/2007, 671–696.

Brusilovsky, P., Maybury, M.T., 2002. From adaptive hypermedia to the adaptive web. Commun. ACM 45, 30–33.

Brusilovsky, P., Rizzo, R., 2003. Using maps and landmarks for navigation between closed and open corpus hyperspace in Web-based education. New Rev. Hypermedia Multimedia 8, 59–82.

Brusilovsky, P., Schwarz, E., Weber, G., 1996. ELM-ART: An intelligent tutoring system on World Wide web. Springer Verlag, pp. 261–269.

Brusilovsky, P., Sosnovsky, S., Shcherbinina, O., 2005a. User Modeling in a Distributed E-Learning Architecture, in: User Modeling 2005. pp. 387–391.

Brusilovsky, P., Sosnovsky, S., Yudelson, M., 2005b. Ontology-based framework for user model interoperability in distributed learning environments. Proceedings of World Conference on E-Learning, E-Learn 2005 2851–2855.

Buneman, P., Davidson, S., Fernandez, M., Suciu, D., 1997. Adding structure to unstructured data. Database Theory—ICDT'97 336–350.

Burke, R., 2002. Hybrid Recommender Systems: Survey and Experiments. User Modeling and User-Adapted Interaction 12, 331–370.

Burke, R.D., Felfernig, A., Gker, M.H., 2011. Recommender systems: An overview. AI Magazine 32, 13–18.

Cain, B., 2007. A review of the mental workload literature. DTIC Document.

Carmagnola, F., Cena, F., 2006. An approach for evaluating User Model Data in an interoperability scenario, in: Proceeding of the 2006 Conference on STAIRS 2006: Proceedings of the Third Starting AI Researchers' Symposium. IOS Press, pp. 251–252.

Carmagnola, F., Cena, F., 2009. User identification for cross-system personalisation. Inf. Sci. 179, 16–32.

Carmagnola, F., Cena, F., Gena, C., 2011. User model interoperability: a survey. User Modeling and User-Adapted Interaction 21, 285–331.

Case, D.O., 2012. Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior. Emerald Group Publishing.

Catledge, L.D., Pitkow, J.E., 1995. Characterizing browsing strategies in the World-Wide web. Computer Networks and ISDN Systems 27, 1065–1073.

Ceri, S., Fraternali, P., Bongio, A., 2000. Web Modeling Language (WebML): a modeling language for designing Web sites. Comput. Netw. 33, 137–157.

Chen, L., Sycara, K., 1998. WebMate: A Personal Agent for Browsing and Searching, in: In Proceedings of the Second International Conference on Autonomous Agents. ACM Press, pp. 132–139.

Chen, Y.-C., Shang, R.A., Kao, C.Y., 2009. The effects of information overload on consumers' subjective state towards buying decision in the internet shopping environment. Electronic Commerce Research and Applications 8, 48–58.

Chi, E.H., Pirolli, P., Chen, K., Pitkow, J., 2001. Using information scent to model user information needs and actions and the Web, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '01. ACM, New York, NY, USA, pp. 490–497.

Claypool, M., Brown, D., Le, P., Waseda, M., 2001. Inferring user interest. IEEE Internet Computing 5, 32 –39.

Collins, A.M., Loftus, E.F., 1975. A spreading-activation theory of semantic processing. Psychological Review 82, 407–428.

Conlan, O., 2004. The Multi-Model, Metadata driven approach to Personalised eLearning Services. Trinity College Dublin.

Conlan, O., Hockemeyer, C., Wade, V., Albert, D., 2003. Metadata Driven Approaches to Facilitate Adaptivity in Personalized eLearning Systems. Journal of the Japanese Society for Information and Systems in Education.

Conlan, Wade, 2004. Evaluation of APeLS–an adaptive eLearning service based on the multi-model, metadata-driven approach, in: Adaptive Hypermedia and Adaptive Web-Based Systems. Springer, pp. 291–295.

Cowie, J., Lehnert, W., 1996. Information extraction. Commun. ACM 39, 80–91.

Cunningham, H., 2002. GATE, a general architecture for text engineering. Computers and the Humanities 36, 223–254.

Cunningham, H., 2005. Information extraction, automatic. Encyclopedia of Language and Linguistics, 665–677.

Dagger, D., O'Connor, A., Lawless, S., Walsh, E., Wade, V.P., 2007. Service-Oriented E-Learning Platforms: From Monolithic Systems to Flexible Services. IEEE Internet Computing 11, 28 –35.

De Bra, P., Houben, G.-J., Wu, H., 1999. AHAM: a Dexter-based reference model for adaptive hypermedia, in: Proceedings of the Tenth ACM Conference on Hypertext and Hypermedia : Returning to Our Diverse Roots: Returning to Our Diverse Roots, HYPERTEXT '99. ACM, New York, NY, USA, pp. 147–156.

Dervin, B., 2012. Sense-Making Studies [WWW Document]. URL http://communication.sbs.ohio-state.edu/sense-making/ (accessed 11.19.12).

Díaz, O., Arellano, C., Iturrioz, J., 2008. Layman tuning of websites: facing change resilience, in: Proceedings of the 17th International Conference on World Wide Web. ACM, pp. 1127–1128.

Dieberger, A., Dourish, P., Höök, K., Resnick, P., Wexelblat, A., 2000. Social navigation: techniques for building more usable systems. interactions 7, 36–45.

Dix, A., 2004. Human-Computer Interaction. Pearson Education.

Dolog, P., Henze, N., Nejdl, W., Sintek, M., 2004. Personalization in distributed e-learning environments, in: WWW Alt. '04: Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters. ACM, New York, NY, USA, pp. 170–179.

Dolog, P., Nejdl, W., 2003. Challenges and Benefits of the Semantic Web for User Modelling.

Edmunds, A., Morris, A., 2000a. The problem of information overload in business organisations: a review of the literature. International Journal of Information Management 20, 17–28.

Edmunds, A., Morris, A., 2000b. The problem of information overload in business organisations: a review of the literature. International Journal of Information Management 20, 17–28.

Edwards, D.M., Hardman, L., 1999. Lost in hyperspace: cognitive mapping and navigation in a hypertext environment 90–105.

Elbassuoni, S., Ramanath, M., Weikum, G., 2011. Query Relaxation for Entity-Relationship Search, in: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., Leenheer, P., Pan, J. (Eds.), The Semanic Web: Research and Applications. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 62–76.

Epstein, P.E., 2009. Beyond information: Exploring patients' preferences. JAMA 302, 195–197.

Feild, H.A., Allan, J., Jones, R., 2010. Predicting searcher frustration, in: Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, Geneva, Switzerland, pp. 34–41.

Feldman, R., Sanger, J., 2006. The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press.

Feuz, M., Fuller, M., Stalder, F., 2011. Personal Web searching in the age of semantic capitalism: Diagnosing the mechanisms of personalisation. First Monday 16, 0–0.

Firmenich, S., Rossi, G., Urbieta, M., Gordillo, S., Challiol, C., Nanard, J., Nanard, M., Araujo, J., 2010. Engineering concern-sensitive navigation structures, concepts, tools and examples. Journal of Web Engineering 9, 157–185.

Fischer, G., 2001. User Modeling in Human–Computer Interaction. User Modeling and User-Adapted Interaction 11, 65–86.

Fleming, J., Koman, R., 1998. Web navigation: designing the user experience. O'reilly Sebastopol, CA.

Gao, M., Liu, K., Wu, Z., 2009. Personalisation in web computing and informatics: Theories, techniques, applications, and future research. Information Systems Frontiers 12, 607–629.

Gauch, S., Speretta, M., Chandramouli, A., Micarelli, A., 2007. User Profiles for Personalized Information Access, in: Brusilovsky, P., Kobsa, A., Nejdl, W. (Eds.), The Adaptive Web, Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 54–89.

Ginige, A., Murugesan, S., 2001. Web engineering: An introduction. Multimedia, IEEE 8, 14–18.

Ginzburg, J., Distante, D., Rossi, G., Urbieta, M., 2009. Oblivious integration of volatile functionality in web application interfaces. Journal of Web Engineering 8, 25.

Google Analytics, 2012. Tracking Code: Basic Configuration - Google Analytics — Google Developers [WWW Document]. URL https://developers.google.com/analytics/devguides/collection/gajs/methods/gaJSApiBasicConfiguration (accessed 8.6.12).

Gruber, T.R., 1993. A translation approach to portable ontology specifications. Knowledge acquisition 5, 199–220.

Guy, I., Carmel, D., 2011. Social recommender systems, in: Proceedings of the 20th International Conference Companion on World Wide Web, WWW '11. ACM, New York, NY, USA, pp. 283–284.

Han, H., Elmasri, R., 2004. Learning Rules for Conceptual Structure on the Web. Journal of Intelligent Information Systems 22, 237–256.

Hauger, D., Paramythis, A., Weibelzahl, S., 2011. Using browser interaction data to determine page reading behavior. User Modeling, Adaption and Personalization 147–158.

Hellerstein, L., Pillaipakkamnatt, K., Raghavan, V., Wilkins, D., 1996. How many queries are needed to learn? Journal of the ACM (JACM) 43, 840–862.

Henze, N., Herrlich, M., 2004. The Personal Reader: A Framework for Enabling Personalization Services on the Semantic Web, in: In Proceedings of the Twelfth GIWorkshop on Adaptation and User Modeling in Interactive Systems (ABIS 04.

Henze, N., Nejdl, W., 2001. Adaptation in Open Corpus Hypermedia. International Journal of Artificial Intelligence in Education 325–350.

Henze, N., Nejdl, W., 2006. Knowledge modeling for open adaptive hypermedia, in: Adaptive Hypermedia and Adaptive Web-based Systems. Springer, pp. 174–183.

Herder, E., Juvina, I., 2004. Discovery of individual user navigation styles.

Herder, E., van Dijk, B., 2005. Site Structure and User Navigation: Models, Measures. Adaptable and Adaptive Hypermedia Systems 19.

Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T., 2004. Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst. 22, 5–53.

Heymann, P., Koutrika, G., Garcia-Molina, H., 2008. Can social bookmarking improve web search?, in: Proceedings of the International Conference on Web Search and Web Data Mining, WSDM '08. ACM, New York, NY, USA, pp. 195–206.

Hiltz, S.R., Turoff, M., 1985. Structuring computer-mediated communication systems to avoid information overload. Commun. ACM 28, 680–689.

Horvitz, E., Breese, J., Heckerman, D., Hovel, D., Rommelse, K., 1998. The lumière project: Bayesian user modeling for inferring the goals and needs of software users, in: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, UAI'98. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 256–265.

Huberman, B.A., Pirolli, P.L., Pitkow, J.E., Lukose, R.M., 1998. Strong regularities in world wide web surfing. Science 280, 95–97.

Hutchinson, A., 1994. Algorithmic learning. Clarendon Press.

Isakowitz, T., Bieber, M., Vitali, F., 1998. Web information systems. Communications of the ACM 41, 78–80.

J. Xiao, Catrambone, R., Stasko, J., 2003. Be Quiet ? Evaluating Proactive and Reactive User Interface Assistants. interactions 383.

Jameson, A., 2009. Adaptive interfaces and agents. Human-Computer Interaction: Design Issues, Solutions, and Applications 105.

Jansen, B.J., Booth, D.L., Spink, A., 2008. Determining the informational, navigational, and transactional intent of Web queries. Information Processing & Management 44, 1251–1266.

Johnson, F., Kumar Gupta, S., 2012. Web Content Mining Techniques: A Survey. International Journal of Computer Applications 47, 44–50.

Juvina, I., Herder, E., 2005. The Impact of Link Suggestions on User Navigation and User Perception, in: Ardissono, L., Brna, P., Mitrovic, A. (Eds.), User Modeling 2005. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 483–492.

Kalbach, J., 2007. Designing Web navigation: Optimizing the user experience. O'Reilly Media, Incorporated.

Kay, J., 2006. Scrutable Adaptation: Because We Can and Must.

Keenoy, K., Levene, M., 2005. Personalisation of Web Search, in: Mobasher, B., Anand, S. (Eds.), Intelligent Techniques for Web Personalization, Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 201–228.

Kellar, M., Watters, C., Shepherd, M., 2006a. The impact of task on the usage of web browser navigation mechanisms, in: Proceedings of Graphics Interface 2006. Canadian Information Processing Society, pp. 235–242.

Kellar, M., Watters, C., Shepherd, M., 2006b. A Goal-based Classification of Web Information Tasks. Proceedings of the American Society for Information Science and Technology 43, 1–22.

Kim, W., 2002. Personalization: Definition, Status, and Challenges Ahead. The Journal of Object Technology 1, 29.

Kirchhoff, M.D., 2013. Enaction: Toward a New Paradigm for Cognitive Science. Philosophical Psychology 26, 163–167.

Knutov, E., De Bra, P., Pechenizkiy, M., 2009. AH 12 years later: a comprehensive survey of adaptive hypermedia methods and techniques. New Review of Hypermedia and Multimedia 15, 5–38.

Kobsa, A., Koenemann, J., Pohl, W., 2001. Personalised hypermedia presentation techniques for improving online customer relationships. The Knowledge Engineering Review 16, 111–155.

Koch, N., Knapp, A., Zhang, G., Baumeister, H., 2008. Uml-based web engineering. Web Engineering: Modelling and Implementing Web Applications 157–191.

Koch, N., Wirsing, M., 2001. Software engineering for adaptive hypermedia applications, in: 8th International Conference on User Modeling, Sonthofen, Germany. Citeseer.

Kohavi, R., Deng, A., Frasca, B., Longbotham, R., Walker, T., Xu, Y., 2012. Trustworthy online controlled experiments: five puzzling outcomes explained, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 786–794.

Koidl, K., Conlan, O., 2008. Engineering Information Systems towards facilitating Scrutable and Configurable Adaptation. Presented at the Fifth International Conference on Adaptive Hypermedia and Adaptive Web-based Systems (AH2008), Springer, Hannover, Germany, pp. 405 – 409.

Koidl, K., Conlan, O., Hampson, Cormac, 2010. Dynamically Adjusting Digital Educational Games Towards Learning Objectives [WWW Document]. URL http://www.tara.tcd.ie/handle/2262/49405 (accessed 11.19.12).

Koidl, K., Conlan, O., Wade, V., 2009. Non-Invasive Adaptation Service for Web-based Content Management Systems, in: International Workshop on Dynamic and Adaptive Hypertext: Generic Frameworks, Approaches and Techniques. Presented at the Hypertext 2009, Torino, Italy.

Komiak, S.Y.X., Benbasat, I., 2006. The effects of personalization and familiarity on trust and adoption of recommendation agents. Mis Quarterly 941–960.

Korpipää, P., Häkkilä, J., Kela, J., Ronkainen, S., Känsälä, I., 2004. Utilising context ontology in mobile device application personalisation, in: Proceedings of the 3rd International Conference on Mobile and Ubiquitous Multimedia, MUM '04. ACM, New York, NY, USA, pp. 133–140.

Korth, H.F., Silberschatz, A., 1997. Database research faces the information explosion. Commun. ACM 40, 139–142.

Kosala, R., Blockeel, H., 2000. Web mining research: a survey. SIGKDD Explor. Newsl. 2, 1–15.

Kumaran, G., Allan, J., 2008. Adapting information retrieval systems to user queries. Information Processing & Management 44, 1838–1862.

Ledford, J.L., Teixeira, J., Tyler, M.E., 2010. Google analytics. Wiley.

Lee, C.H., Cranage, D.A., 2011. Personalisation–privacy paradox: The effects of personalisation and privacy assurance on customer responses to travel Web sites. Tourism Management 32, 987–994.

Licklider, J.C.R., Taylor, R.W., others, 1968. The computer as a communication device. Science and technology 76, 2.

Lieberman, H., 1995. Letizia: An Agent That Assists Web Browsing, in: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE. pp. 924–929.

Lievrouw, L.A., 1996. The information society: A study of continuity and change. Journal of the American Society for Information Science 47, 250–251.

Liewehr, S., Laplante, M., 2011. Understanding Best Practices for Profiling, Personalizing, and Targeting Next Generation Engagement.

Linden, G., Smith, B., York, J., 2003. Amazon.com recommendations: item-to-item collaborative filtering. IEEE Internet Computing 7, 76 – 80.

Mackay, D.M., 1960. What makes the question, The Listener.

MacRae, D.G., Wolff, K.H., 1951. Georg Simmel: The Sociology of Georg Simmel. The Economic Journal 61, 163.

Maedche, A., Staab, S., 2001. Ontology learning for the semantic web. Intelligent Systems, IEEE 16, 72–79.

Meadow, C.T., 1992. Text Information Retrieval Systems. Academic Press, Inc., Orlando, FL, USA.

Miculan, M., Urban, C., 2011. Formal analysis of Facebook Connect single sign-on authentication protocol, in: SOFSEM. pp. 22–28.

Mitra, M., Singhal, A., Buckley, C., 1998. Improving automatic query expansion, in: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, pp. 206–214.

Mobasher, B., Cooley, R., Srivastava, J., 2000. Automatic personalization based on Web usage mining. Commun. ACM 43, 142–151.

Montebello, M., Gray, W.A., Hurley, S., 1998. An Evolvable Personal Advisor to Optimize Internet Search Technologies, in: DEXA. pp. 531–540.

Morita, M., Shinoda, Y., 1994. Information filtering based on user behavior analysis and best match text retrieval, in: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '94. Springer-Verlag New York, Inc., New York, NY, USA, pp. 272–281.

Moukas, A., Maes, P., 1998. Amalthaea: An Evolving Multi-Agent Information Filtering and Discovery System for the WWW. Autonomous Agents and Multi-Agent Systems 1, 59–88.

Mulvenna, M.D., Anand, S.S., Büchner, A.G., 2000. Personalization on the Net using Web mining: introduction. Commun. ACM 43, 122–125.

Nanas, N., Roeck, A., Vavalis, M., 2009. What Happened to Content-Based Information Filtering?, in: Azzopardi, L., Kazai, G., Robertson, S., Rüger, S., Shokouhi, M., Song, D., Yilmaz, E. (Eds.), Advances in Information Retrieval Theory. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 249–256.

Noll, M.G., Meinel, C., 2007. Web Search Personalization Via Social Bookmarking and Tagging, in: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (Eds.), The Semantic Web. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 367–380.

O'Callaghan, T., 2011. Eli Pariser: The dark side of web personalisation. The New Scientist 211, 23.

Oard, D.W., Kim, J., 2001. Modeling Information Content Using Observable Behavior.

Oberle, D., Berendt, B., Hotho, A., Gonzalez, J., 2003. Conceptual User Tracking, in: Menasalvas, E., Segovia, J., Szczepaniak, P. (Eds.), Advances in Web Intelligence, Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 955–955.

Olston, C., Chi, E.H., 2003. ScentTrails: Integrating browsing and searching on the Web. ACM Trans. Comput.-Hum. Interact. 10, 177–197.

Paramythis, A., Weibelzahl, S., Masthoff, J., 2010. Layered evaluation of interactive adaptive systems: framework and formative methods. User Modeling and User-Adapted Interaction 20, 383–453.

Parikh, R., Movassate, M., 2009. Sentiment analysis of user-generated twitter updates using various classification techniques. CS224N Final Report.

Pariser, E., 2011. The filter bubble: What the Internet is hiding from you. Penguin Press HC.

Park, J., Kim, J., 2000. Effects of contextual navigation aids on browsing diverse Web systems, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, The Hague, The Netherlands, pp. 257–264.

Parker, E.B., Paisley, W.J., 1966. Research for psychologists at the interface of the scientist and his information system. American Psychologist 21, 1061.

Pastor, O., Fons, J., Pelechano, V., Abrahão, S., 2006. Conceptual modelling of web applications: the OOWS approach. Web Engineering 277–302.

Pazzani, M., Billsus, D., 1997. Learning and Revising User Profiles: The Identification of Interesting Web Sites. Machine Learning 27, 313–331.

Pazzani, M., Billsus, D., 2007. Content-Based Recommendation Systems, in: Brusilovsky, P., Kobsa, A., Nejdl, W. (Eds.), The Adaptive Web, Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 325–341.

Perugini, S., Ramakrishnan, N., 2003. Personalizing Web sites with mixed-initiative interaction. IT Professional 5, 9 – 15.

Pirolli, P., Card, S., 1999. Information foraging. Psychological Review 106, 643–675.

Pol, K., Patil, N., Patankar, S., Das, C., 2008. A Survey on Web Content Mining and Extraction of Structured and Semistructured Data, in: First International Conference on Emerging Trends in Engineering and Technology, 2008. ICETET '08. Presented at the First International Conference on Emerging Trends in Engineering and Technology, 2008. ICETET '08, pp. 543 –546.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J., 1994. GroupLens: an open architecture for collaborative filtering of netnews, in: Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work, CSCW '94. ACM, New York, NY, USA, pp. 175–186.

Riloff, E., Lehnert, W., 1994. Information extraction as a basis for high-precision text classification. ACM Trans. Inf. Syst. 12, 296–333.

Rizzo, G., 2011. Nerd: Evaluating named entity recognition tools in the web of data 1–16.

Rossi, G., Schwabe, D., 2008. Modeling and implementing web applications with OOHDM. Web Engineering: Modelling and Implementing Web Applications 109–155.

Rudd, M.J., Rudd, J., 1986. The Impact of the Information Explosion on Library Users: Overload or Opportunity? Journal of Academic Librarianship 12, 304–06.

Sah, M., Wade, V., 2010. Automatic metadata extraction from multilingual enterprise content, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management. ACM, Toronto, ON, Canada, pp. 1665–1668.

Salton, G., McGill, M.J., 1986. Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York, NY, USA.

Schafer, J., Frankowski, D., Herlocker, J., Sen, S., 2007. Collaborative Filtering Recommender Systems, in: Brusilovsky, P., Kobsa, A., Nejdl, W. (Eds.), The Adaptive Web, Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 291–324.

Schafer, J.B., Konstan, J., Riedi, J., 1999. Recommender systems in e-commerce, in: Proceedings of the 1st ACM Conference on Electronic Commerce, EC '99. ACM, New York, NY, USA, pp. 158–166.

Schafer, J.B., Konstan, J.A., Riedl, J., 2001. E-Commerce Recommendation Applications. Data Mining and Knowledge Discovery 5, 115–153.

Schafer, J.B., Konstan, J.A., Riedl, J., 2002. Meta-recommendation systems: user-controlled integration of diverse recommendations, in: Proceedings of the Eleventh International Conference on Information and Knowledge Management. ACM, pp. 43–51.

Schiaffino, S., Amandi, A., 2004. User – interface agent interaction: personalization issues. International Journal of Human-Computer Studies 60, 129–148.

Schwabe, D., Esmeraldo, L., Rossi, G., Lyardet, F., 2001. Engineering Web applications for reuse. IEEE MultiMedia 8, 20 –31.

Shahabi, C., Chen, Y.-S., 2003. Web information personalization: Challenges and approaches. Databases in Networked Information Systems 5–15.

Sifry's Alerts: State of the Blogosphere, April 2006 Part 1: On Blogosphere Growth [WWW Document], n.d. URL http://www.sifry.com/alerts/archives/000432.html (accessed 4.22.13).

Smith, A.S.G., Blandford, A., 2003. MLTutor: An Application of Machine Learning Algorithms for an Adaptive Web-based Information System. International Journal of Artificial Intelligence in Education 13, 235–261.

Smith, P.A., 1996. Towards a practical measure of hypertext usability. Interacting with Computers 8, 365–381.

Sorensen, H., McElligott, M., 1995. PSUN: a profiling system for Usenet news, in: Proceedings of CIKM. Citeseer, pp. 1–2.

Speier, C., Valacich, J.S., Vessey, I., 1999. The Influence of Task Interruption on Individual Decision Making: An Information Overload Perspective. Decision Sciences 30, 337–360.

Speretta, M., Gauch, S., 2005. Personalized search based on user search histories, in: The 2005 IEEE/WIC/ACM International Conference on Web Intelligence, 2005. Proceedings. Presented at the The 2005 IEEE/WIC/ACM International Conference on Web Intelligence, 2005. Proceedings, pp. 622 – 628.

Stadnyk, I., Kass, R., 1992. Modeling users' interests in information filters. Commun. ACM 35, 49–50.

Staikopoulos, A., O'Keeffe, I., Rafter, R., Walsh, E., Yousuf, B., Conlan, O., Wade, V., 2012. AMASE: A Framework for Composing Adaptive and Personalised Learning Activities on the Web, in: Popescu, E., Li, Q., Klamma, R., Leung, H., Specht, M. (Eds.), Advances in Web-Based Learning - ICWL 2012. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 190–199.

Stanley, A.J., Clipsham, P.S., 1997. Information overload-myth or reality?, in: IT Strategies for Information Overload (Digest No: 1997/340), IEE Colloquium On. pp. 1/1 –1/4.

Steichen, B., Lawless, S., O'Connor, A., Wade, V., 2009. Dynamic hypertext generation for reusing open corpus content, in: Proceedings of the 20th ACM Conference on Hypertext and Hypermedia. pp. 119–128.

Stewart, A., Diaz-Aviles, E., Nejdl, W., Marinho, L.B., Nanopoulos, A., Schmidt-Thieme, L., 2009. Cross-tagging for personalized open social networking, in: Proceedings of the 20th ACM Conference on Hypertext and Hypermedia. ACM, Torino, Italy, pp. 271–278.

Sweeney, L., Zayatz, P., Doyle, J., 2001. Information Explosion. Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies. Urban Institute.

Sweller, J., 1988. Cognitive load during problem solving: Effects on learning. Cognitive science 12, 257–285.

Tan, A., 1999. Text Mining: The state of the art and the challenges, in: In Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases. pp. 65–70.

Taylor, R.S., 1967. Question-Negotiation an Information- seeking in libraries. DTIC Document.

Teevan, J., Alvarado, C., Ackerman, M.S., Karger, D.R., 2004. The perfect search engine is not enough: a study of orienteering behavior in directed search, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '04. ACM, New York, NY, USA, pp. 415–422.

Terai, H., Saito, H., Egusa, Y., Takaku, M., Miwa, M., Kando, N., 2008. Differences between informational and transactional tasks in information seeking on the web, in: Proceedings of the Second International Symposium on Information Interaction in Context, IIiX '08. ACM, New York, NY, USA, pp. 152–159.

Uther, J., Kay, J., 2003. VlUM, a web-based visualisation of large user models. User Modeling 2003 145–146.

Uther, J.B., Uther, C.J.B., 2001. On the Visualisation of Large User Model in Web Based Systems.

Webb, G.I., Pazzani, M.J., Billsus, D., 2001. Machine Learning for User Modeling. User Modeling and User-Adapted Interaction 11, 19–29.

White, R.W., Bailey, P., Chen, L., 2009. Predicting user interests from contextual information, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09. ACM, New York, NY, USA, pp. 363–370.

White, R.W., Ruthven, I., Jose, J.M., 2002. Finding relevant documents using top ranking sentences: an evaluation of two alternative schemes, in: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '02. ACM, New York, NY, USA, pp. 57–64.

Widyantoro, D.H., Yin, J., Nasr, M.S.E., Seif, M., Nasr, E., Yang, L., Zacchi, A., Yen, J., 1999. Alipes: A Swift Messenger in Cyberspace, in: In Proc. AAAI Spring Symposium on Intelligent Agents in Cyberspace (Palo Alto. AAAI Press, pp. 62–67.

Witten, I.H., Frank, E., 2005. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Witten, I.H., Frank, E., Trigg, L.E., Hall, M.A., Holmes, G., Cunningham, S.J., 1999. Weka: Practical machine learning tools and techniques with Java implementations.

Wood, J., Wright, H., Brodie, K., 1997. Collaborative visualization, in: Visualization'97., Proceedings. IEEE, pp. 253–259.

Xu, J., Croft, W.B., 1996. Query expansion using local and global document analysis, in: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, pp. 4–11.

Yinghui, Y., 2010. Web user behavioral profiling for user identification. Decision Support Systems 49, 261–271.

Yudelson, M., 2010. Providing Service-Based Personalization In An adaptive Hypermedia System, PHD Thesis. University of Pittsburgh 2010.

Zhou, X., Gaugaz, J., Balke, W.T., Nejdl, W., 2007. Query relaxation using malleable schemas, in: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data. ACM, pp. 545–556.

Žumer, M., 2012. Interactive Information Seeking, Behaviour and Retrieval. Program: electronic library and information systems 46, 277–278.

Appendices

Appendix I – State of the Art Survey of Web-based Content Management Systems

Introduction

In recent years Web based Content Management Systems (WCMS) concentrating on the creation and management of web content such as Wikipedia, Drupal, WordPress etc. have become an important part of the overall Internet experience. A vast community of Internet users and developers refer to WCMS in order to interact, publish, share experiences, knowledge and information. From a user's point of view these systems allow an easy creation and management of web content without the need of using Internet mark-up or programming languages. Nevertheless the development of easy to use WCMS interfaces and the management of vast amounts of web content pose numerous challenges to the developers. Especially the integration of additional features and interactions within the core architecture of the WCMS can trigger unexpected system behaviours. Open Source projects, most famously MediaWiki and Drupal, provide extension possibilities without the need of altering the core of the WCMS. Depending on the WCMS extensions can range from entire framework plug-ins to simple API's.

The objective of this survey is to investigate the possibilities for integrating third-party Adaptive WCMS by using existing extension possibilities and avoiding and avoiding changes in the WCMS core. For this the survey will focus on architectural issue regarding the integration of additional modules, API usage, the availability of user and content/domain related models and other technical issues. This survey is part of a PhD research project developing such a third party Adaptive Service for WCMS. The adaptive interventions of this third party adaptive Service are to be non-intrusive on the users experience with the used WCMS and range from simple adaptive presentation techniques like altering the navigation possibilities to more complex content recommendations based on user preferences and interests.

The survey is divided into the following main chapters: The survey outline classifying WCMS and providing a general technical overview. This is followed by the main survey comparing two successful Open Source WCMS projects, namely Drupal and MediaWiki, and finally a summarisation towards third party Adaptive Service integration in Open Source WCMS.

Survey Outline

Classifying Web Content Management Systems

Generally WCMS can range from traditional electronic document management to complex organisation wide Customer Relationship Management (CRM). Most vendors including Open Source communities though give the impression that the term WCMS refers to a specific product or deployment. The most distinctive difference between a WCMS and a traditional CMS is the accessibility via common Internet tools and applications. Based on this distinction WCMS are often referred to as a Web Applications.[129] Taking a broader view on WCMS though shows that a WCMS refers to a whole range of different applications and functionalities with the main commonality of providing effective and accessible tools for publishing and organising web based content. A notable journal publication by Goodwin and Vidgen defines a WCMS as the following:

'*an organizational process, aided by software tools, for the management of heterogeneous content on the web, encompassing a life cycle that runs from creation to destruction*' (Vidgen, 2002)

Furthermore the term WCMS mostly refer to Open Source based content management system and WCMS with a commercial background are referred to as Enterprise Content Management Systems (ECM) providing the same or in some cases even more functionalities based on being embedded in cooperate frameworks like Adobe Systems Contribute Publishing Server (CPS) or Microsoft Office SharePoint Server Services. In the following the term WCMS is referred to as Open Source WCMS.

In recent years numerous WCMS have appeared and have become an enormously popular application domain in day to day Internet activities. Systems like Wikipedia, MySpace, Twitter, WordPress have become well known to most Internet users. WCMS include applications that range from traditional print domains like photo galleries and news articles to discussion boards and the creation and management of entire community and corporate web sites including interactive product catalogues and e-commerce functionalities.

One of the main reasons why WCMS are enjoying such popularity is based on the fact that the creation and deployment of web content is simple and easy accessible for non-technical Internet users. This allows users to easily access edit and manage content without needing to know anything about implementation languages or the back end architecture.

---

[129] For a basic overview of WCMS and brief descriptions please note http://en.wikipedia.org/wiki/Web_content_management_system

General technical overview

In order to explore the wide ranging technical approaches used for the design of WCMS it seems most appropriate to firstly introduce a high level classification based on the content itself. Based on content processing the WCMS are generally classified into three different types[130]:

Online Processing is based on request time generation of content. Depending on the WCMS the generation mostly depends on different caching models. Known online processing WCMS are Drupal, Joomla!, TYPO3 and Mambo.

Offline Processing allows the user to pre-process all content mostly with the aid of different offline templates. The content is not generated on the fly at request time and therefore requires no major request time server side actions.

Hybrid Systems combine offline and online processing by not only sending static content to the client but also executable code.

Some WCMS provide additional content management workflows to ensure higher quality content by managing the life cycle of a document. E.g. Drupal provides a workflow management module which can be added to the core implementation. MediaWiki on the other hand does not provide any document workflow management, but effectively provides revision possibilities on individual articles. Even though most WCMS still focus on the content it can be noted that some WCMS are slowly evolving into *Web Application Frameworks*. E.g. Drupal only provides a very minimal core implementation allowing easy extension with pluggable modules. Other WCMS specialise on a specific domain like WordPress for blogging or Joomla! For template based web page creation.

In the following a general technical overview of specific WCMS functional and non-functional components is given. The purpose is to provide a simple overview and introduction into necessary components and technologies used in WCMS with a focus on third party adaptive interventions.

Functional components

User behaviour is the most essential information necessary in order to provide adaptive functionalities. In regards to simply browsing content the user behaviour allows the tracking of interests and preferences. More active tasks like creating content or adding comments and ratings

[130] Please note http://www.contenthere.net/2007/06/cms-deployment-patterns.html for more information on the design of content deployment within WCMS.

can also be seen as indications for specific interests helpful for adaptation. Tracked user behaviour needs to be saved in a user model that is constantly updated. Unfortunately currently no WCMS provides a rich user tracking or user modelling functionality. Most commonly WCMS provide a log in possibility for identification and a related preference list which is managed by the user and not by the system based on behaviour and interests.

Document Management relates to a whole range of different functions related to the management of the content within a WCMS. This especially involves the workflow of a document ranging from the creation and storage of a document to the revision management. Depending on the broadness of WCMS there are different approaches. Drupal for instance provides a Document Management Module including the most important features and functions. MediaWiki on the other hand uses an online model with additional concerns such as caching.

Interfaces and Modules have become important components within WCMS. Depending on the nature of the WCMS being either commercial or open source the API can be used to access WCMS with third party web applications. On the other hand WCMS can be extended with modules which in the case of Drupal, as a WCMS which heavily relies on interfaces and modules, can be plugged into the core implementation. New modules are continuously developed extended based on Drupals huge Open Source developing community. MediaWiki on the other hand only provides limited extensions based on modules although still enjoys immense popularity throughout its developer community.

Non-Functional components

Coupled and Decoupled WCMS refers to the coupling between the content location and the Webserver deploying the content. In a coupled scenario the architecture is easier to manage although runtime issues get more relevant because of the missing separation of concerns. On the other hand coupled WCMS provide more possibilities for personalising content. Most prominent examples for coupled WCMS are Drupal and Plone for decoupled WCMS RedDot.

Caching has turned out to be one of the most important performance issue related to response time. Especially for MediaWiki caching has become a curtail element. For local WCMS though with smaller communities it has to be noted that response time is important, but is not always seen as crucial.

Security in WCMS can range from very simple login functionalities like login history to providing more complex security features like Secure Socket Layer (SSL) connections and certificate

management. For this research especially the authentication of the user is essential. In addition security features such as granular privileges related to altering of specific content can be useful.

Web Template Systems also referred to as Web Template Engine has the function to provide a user friendly approach for the publishing of web content on the WCMS. For this the system mostly provides a set of different templates. Typically these templates follow a specific branding providing consistency in the look and feel of a WCMS. Template approaches are particularly successful based on the time that can be saved in developing entire web sites. WCMS such as Joomla! or Mambo follows this approach. MediaWiki on the other hand provides an extended template system which can be extended and accessed via MediaWiki's API.

Web Services add additional external features to a WCMS. Web service integration is mostly integrated in the front end directly exposing the services features. The most prominent example is the integration of the Google API allowing an added search bar in the WCMS integrating with the Google servers.

Comparison of Web Content Management Systems

In this chapter firstly the main evaluation criteria for comparing WCMS are introduced and discussed. The evaluation criteria chosen based on the goal of the related PhD research. Finally two prominent WCMS are evaluated based on the initial criteria. These are Drupal and MediaWiki. These systems were not only selected based on their popularity, but on their possibility of extending basic functionalities towards third-party Adaptive Service integration.

Evaluation Criteria

In the technical overview chapter different general technical and non-technical components necessary for a WCMS to function effectively were described. In this chapter specific WCMS criteria in relation to the research goal are discussed.

Application and Usage

This chapter describes the nature of the WCMS. This includes the usage and the typical user groups. Depending on the WCMS also the development society and the outlook/strategy of the WCMS will be discussed.

Architecture

The architecture of a WCMS includes all major technical components important to the specific WCMS. These include the application server, web server, databases, design languages, license issues and the template engines.

Content Management

The content and its management is the core function of a WCMS. Content related issues include the content types, content format and standards, possible export and import interfaces.

Semantics

Semantic feature are the foundation for any adaptive functionality within a WCMS. Especially content/domain models like Ontologies are essential elements. In addition the WCMS needs to provide basic user modelling or a user profile to trigger user interests and match these with related content.

Adaptivity

Adaptive feature refer to the usage of semantic feature to provide adaptivity within the WCMS. This can include simple adaptation presentation like colours and template setting or more complex adaptive navigation like link management or content recommendations.

Policies and Workflows

In order to alter different elements within a WCMS internal policies and workflow related to the changes have to be considered. These can range from simple content location and retrieval mechanisms to complex content management and access policies.

Localisation

Most WCMS especially in the public domain like MediaWiki and Drupal are available in many different languages. Not only by localising the interface and the content but also including entire development communities by exposing localised interfaces and tools. MediaWiki is the most dominant example in which the MediaWiki implementation Wikipedia is used in over 100 languages. It has to be noted though that the articles are not translations but unique articles written in the appropriate language.

Web Service and third part API integration

Popular Open Source WCMS projects mostly expose an API, although API functionalities are mostly limited and only allow the querying of information stored in the WCMS. Extended API which also allows the altering of the design by adding additional elements into the look and feel of the WCMS are rare and very difficult to implement. Advanced extensions make the usage of Web Services necessary allowing a more complex interaction between third party client applications and the WCMS.

Future developments and outlook

Many WCMS are slowly but surely moving toward a Content Management Framework with Drupal being the most dominant example. As mentioned before other WCMS evolve and specialize into more effective WCMS for specific domains like blogging or discussion boards for tech support. Generally it is difficult to grasp the future development and outlook of open source WCMS based on its vast amount of developers and contributors. The focus of this chapter therefore will be on the development of extensions and possible interactions with third party applications.

Drupal

Application and Usage

Drupals main community web page refers to Drupal as free software package which allows an individual or a community of users to easily publish, manage and organize a wide variety of content on a website.[131] Based on the wide possibilities Drupal offers, it has to be noted that Drupal is generally seen as more than a WCMS. Drupal caters a wide variety of web application e.g. Community Web Portals, Intranet Applications, Personal Web Sites, Blogging interfaces, E-commerce applications, Social Networking Applications and many more. In order to use Drupal as WCMS various modules can be used and plugged into the basic core implementation. Drupal is supported by a big development community constantly extending the core architecture with extensions and pluggable models.

Architecture

As mentioned in the application and usage section Drupal is not strictly a WCMS with a specific application area. Drupal is often referred to as Content Management Framework (Drupal, 2009).

[131] Please note http://drupal.org/about for more general information on Dural.

Therefore Drupal does not specialise in one specific application area like blogging or social networking. By providing an extendible core implementation that can plug different modules depending on the application area, it provides an architecture which caters for many different types of application areas. This high level of flexibility and abstraction though comes with the cost that developers need to have good knowledge of the core elements in order to extend them. Following images provide a brief overview of the architecture and the core elements of Drupal. Following figure 1 indicates the separation of concerns within the architecture by providing different levels that can be accessed individually. All content is managed on the first layer as nodes. Models on the second layer can be plugged accessing and managing the content bellow. The three layers above perform necessary elements such as User Interface, User Permissions and different templates for the look and feel of the front end.

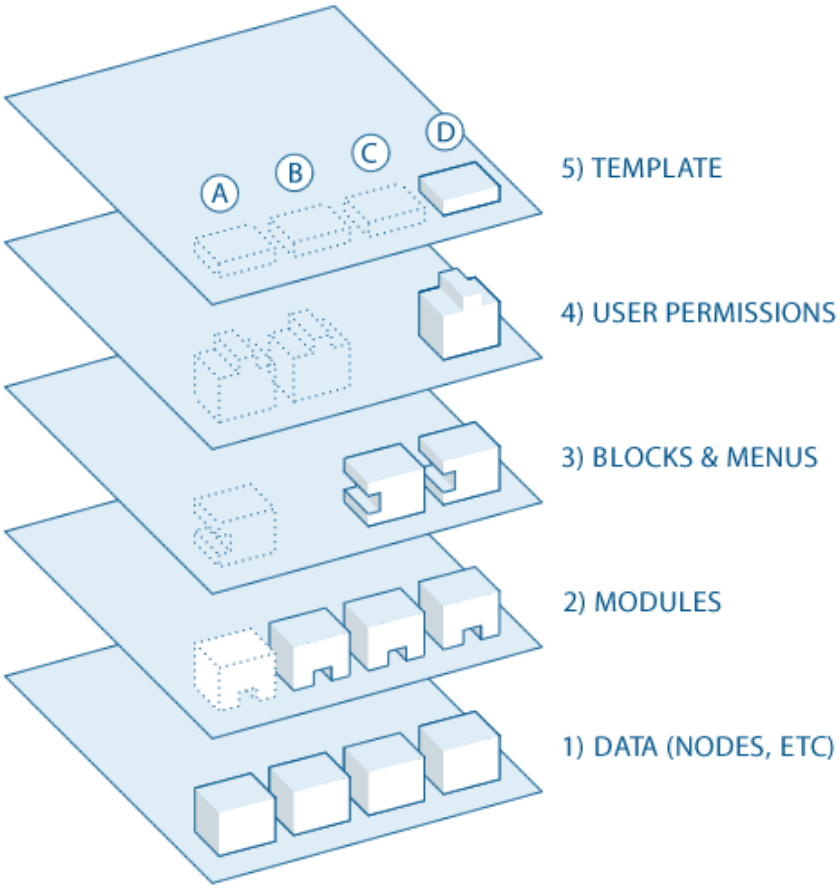


Figure 58 Drupals Layer Architecture (Drupal, 2009)

Drupal is an Open Source project under GPL allowing the community to further develop and extend the main system. Drupals technology stack is based on the LAMP (Linux/Apache/MySQL/PHP)

solution stack described further down. For an overview of the technology stack please note Figure 59:

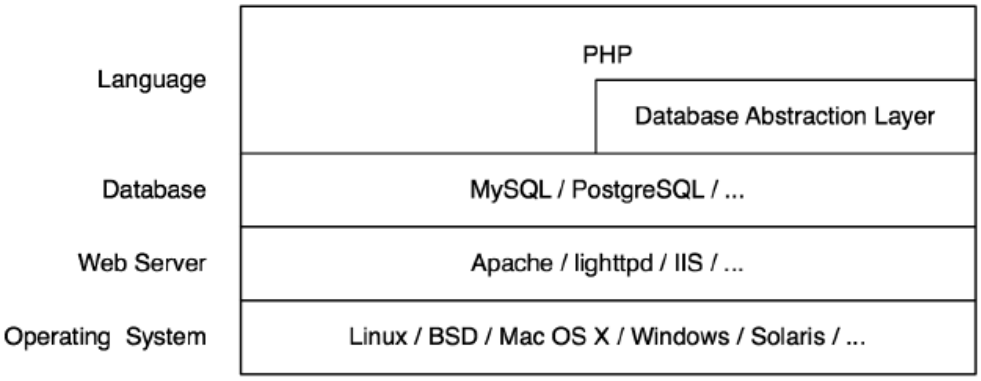


Figure 59 Illustrating Drupals Technology Stack (vanDyk, 2008)

As indicated in Figure 58 the module layer of Drupal is an essential part of the overall Drupal architecture. Drupals core illustrated in Figure 60 enables the plugging of modules of which only a small number is necessary for a basic.

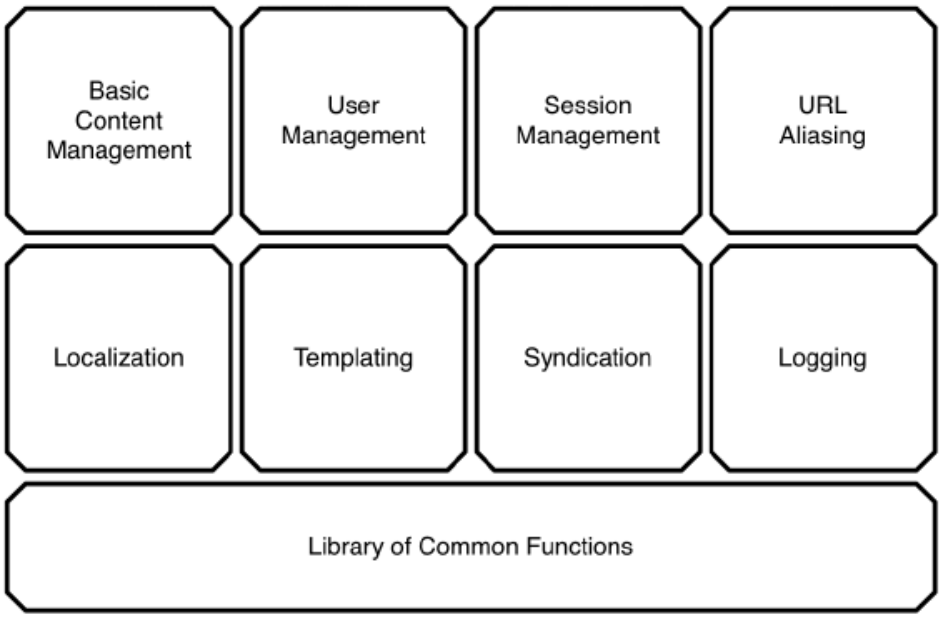


Figure 60 An Overview of Drupals core (vanDyk, 2008)

Content Management

All content in Drupal is organised as nodes. These nodes are managed based on ids although a node does not have to be one specific page or piece of content. A node can be anything that can be stored on a database and which can be integrated as part of web content e.g. a blog, forum entries, forms etc. A key element in the content management of Drupal is the *Content Construction Kit (CCK)*. It allows the creation of different types of content for which the structure can be composed very flexible.

Beside the CCK Drupal also contains a Views module. This model which has to be added to the core allows the flexible management of content access and presentation.

A further interesting module in the context of content management is the Taxonomy module. This module allows the organisation of different taxonomies based on term vocabularies.

Semantics

The Drupal community is involved in creating a Web3.0 extension for Drupal. This extension is based on the previously mentioned node concept and the Taxonomy module. Please note Figure 61 for the initial architecture of Drupals Semantic Web approach. The basic idea is to provide Semantic Search functionalities which can be accessed via an API. For this additional information about the back end content stored on a RDBMS is stored in an ARC RDF format for effective querying. Unfortunately it seems not entirely clear what the additional information will be or where it comes from.

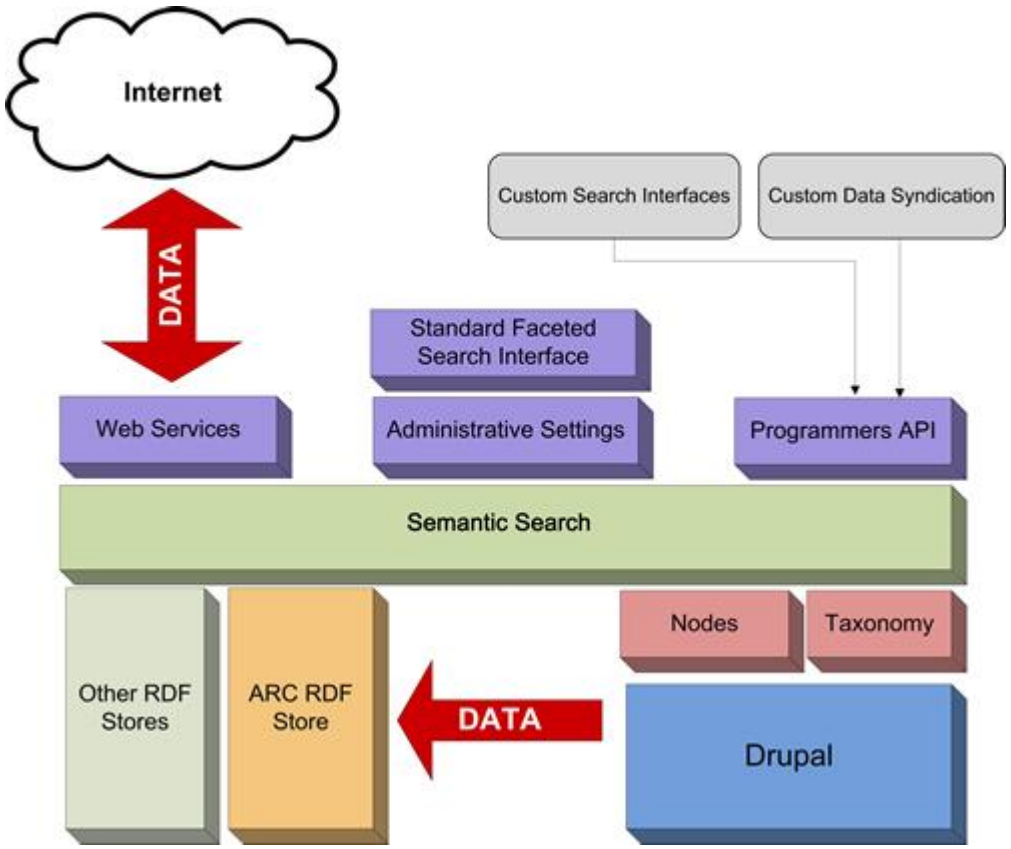


Figure 61 Drupals Semantic Web Architecture (http://semanticsearch.org/)

Finally the Relationship Module allowing the management of relationships between nodes should be mentioned. The relationships created not only indicate a relationship, but also the kind of relationship driving the taxonomy approach towards an ontological based approach.

Adaptivity

An important element to enable adaptive features is the possibility of building user models or user profiles. User models can range from static user created profiles to dynamic system created and updated user models which are based on user activities and behaviour. Drupal provides a customized user profile interface for the user to indicate specific preferences. This information can be used to trigger matches between specific content nodes within the WCMS based on interests from the user profile. In addition the user profile in Drupal can be designed very flexible to cater for different profile sections e.g. Business, Community activities, others. Unfortunately more advanced user modelling feature triggering user activities has not been developed yet.

As second important element for adaptive features within a WCMS is to provide content/domain models to which the interests stated in the user models can be matched. Drupal provides a content management module which enables the indexing and matching of content towards user profile

settings. This feature can be accomplished by using Drupals Taxonomy module which provides term based vocabularies assigned to different content nodes. These taxonomies can be matched with user profile.

Related to the need of content/domain modules for effective adaptation Drupal developers are engaged in several development projects developing ontology based content extension for adaptation needs. One specific project to highlight is the *Semantically-Interlinked Online Community (SIOC)* [132] project which is the providing a social activity based RDF Ontology for the representation social websites in online community based on Drupal. This approach conjuncts with the *Friends-of-a-friend (FOAF)*[133] project which provides vocabularies for user profile matching.

Finally the *MySite* module provides similar functionalities as *iGoogle* or *MyYahoo!*. This module applies to the content available within the installed Drupal WCMS, but can also plug in different web elements like RSS or API based application to enrich the experience.

Policies and Workflows

The main policies in place are related to how to integrate third party code as module or API access based services in Drupal.

MediaWiki

Application and Usage

MediaWiki is known as the main Wiki engine developed originally for the well-known worldwide collaborate free online Encyclopaedia *Wikipedia*. It is an Open Source project and is hosted and managed by the *Wikimedia Foundation*. The term *WikiMedia,* which is easily to be confused with MediaWiki, refers to a whole group of Open Source projects including Wikipedia, but also *Wiktionary*, *Wikisource*, *Wikibooks* and other *Wiki* based projects. Finally the term *Wiki* refers to all MediaWiki related Open Source projects in which users collaborate to freely create, edit, access and publishing knowledge. Recently Wiki's have moved away from being a purely encyclopaedia driven WCMS to a general Knowledge Management System used in communities, ranging from Open Web Communities to internal Enterprise knowledge systems.

[132] Please note http://sioc-project.org/ for further information.

[133] Please note http://www.foaf-project.org/ for further information.

MediaWiki with its main project Wikipedia is online since 2001 and stores over 12 million collaboratively created articles in over 100 different languages and is seen to date as the most popular reference source on the Internet. Based on its huge collaborating community ensuring fast updates in the knowledge base Wikipedia enjoys overwhelming popularity. (http://en.wikipedia.org/wiki/Wikipedia)

The main critic against Wikipedia though is the unreliability of the source. Wikipedia being successful because of its huge voluntary creating, editing and in some case article defending community lacks credentials, accuracy and objectivity in relation to the source of knowledge leading in some cases to a distorted consensus and subjective opinion driven articles.

Architecture

The MediaWiki architecture is highly scalable with strong emphasis on caching and server load balancing to enable accessibility for a high number of simultaneously accessing users. It is therefore suitable to run on large server farms. MediaWiki uses PHP for processing and displaying of web content which is stored in a MySQL database. Like Drupal the entire solution stack is based on a 'LAMP' architecture[134]. For searching and retrieving MediaWiki uses a Lucene.

The architecture of MediaWiki provides a simple way of adding useful functionality to the implementation via the *extensions* feature. These are based on simple scripts adding different plugins like Adobe Flash, Video streaming, RSS feeds, ratings and API based third party accessibility. Recently more enhanced extensions were introduced adding semantic operations like adding semantic structure and relationships to different articles within the content of MediaWiki.[135]

Content Management

The most significant element of MediaWiki is the easy creation of web content. This is done via *wikitext* enabling the adding of formatting and linkage without any knowledge of HTML, XHTML or other Internet mark-up languages. E.g. adding links in an article is done by placing the linkable keyword within two square brackets. The important part hereby is that the keyword within the brackets has to relate to the exact title within the MediaWiki database. This avoids displaying wrong

[134] A LAMP architecture is often used in Open Source projects and refers to **L**inux operating system, **A**pache Web Server, **M**ySQL database and **P**HP (alternatively Perl or Python) programming language.

[135] A complete list of official MediaWiki extensions can be found under http://www.mediawiki.org/wiki/Extension_Matrix

keywords in published articles. In addition also namespaces are possible to allow further distinctions.

A further important content management feature in MediaWiki is the usage of *Portals*. Like the main page of every MediaWiki based system it allows the definition of a specific template related to a specific topic. Instead of using the main page the user can navigate directly to a MediaWiki Portal e.g. 'Car engines' in which only articles related to car engines will be displayed. The layout portal can be organised in different ways allowing different portals to have different designs and layouts depending on the related topic.

In relation to the navigation possibilities MediaWiki enables *Navigation Templates* or *Topic Boxes* allowing the adding links within a defined area anywhere in the article or the overlaying template.

As mentioned above the publishing of articles on MediaWiki is kept as simple as possible, making the creation of an article without any knowledge of web programming languages easy to do. Articles are stored on the related database system and managed via a revision system ensuring that articles can be revised in the case of vandalism or miss information.

MediaWiki features a range of different content types and classifies different usages in reference lists.[136]

Semantics

In order to provide a semantic structure in MediaWiki extensions can be used enabling the further structuring of content and the creation of relationship between different articles. The *MediaWiki Semantic extension* provides different properties which can be assigned to different articles within the MediaWiki repository. This allows articles to be combined and mixed by making them members of specific categories. This rich metadata allows numerous semantic possibilities enhancing not only the query/search of articles, but also the quality of data by avoiding redundant articles.

Following architecture for the Semantic MediaWiki is proposed by Kroetzsch and Vrandecic:

[136] A full list and explanation of all MediWiki content types and references can be found under http://en.wikipedia.org/wiki/Portal:Contents

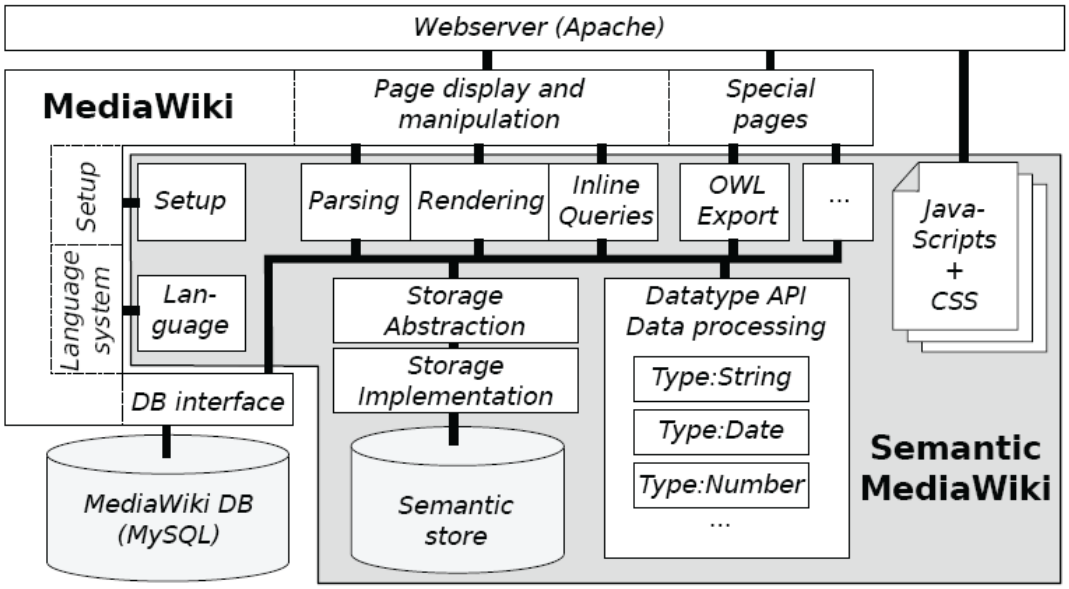


Figure 62 Sematic MediaWiki (Volkel, 2006)

The MediaWiki semantic approach would allow the exporting of MediaWiki Ontologies currently only possible by using the link structure and the article titles without taking further relations into accounts. Different research and commercial projects created RDF representations of different MediaWiki projects at different points of time.[137] The most common approach to create MediaWiki Ontology like structure is to export the title relationships of the published articles in the MediaWiki WCMS.

Adaptivity

As described in the introduction any system or entire WCMS providing adaptive features has to enable a user model or profile and a related domain or content model in order to match interests or queries. MediaWiki's Semantic MediaWiki project will eventually provide the necessary content model to ensure effective adaptation in the navigation and retrieval of the content. Even simple tropic maps including titles and related articles can be used for user model matching. Adaptive presentation though has to be confined in the boundaries of the template structure mentioned above. MediaWiki with its template and portal design therefore potentially can provide both adaptive presentation and adaptive navigation.

[137] Please note the Wikipedia³ (http://labs.systemone.at/wikipedia3) dataset based on a WikiMedia dump in English including approx. 47 million RDF triples. Alternatively a project called Wikitology is trying to create a large Ontology over MediaWiki datasets (http://ebiquity.umbc.edu/resource/html/id/245/Wikitology-Wikipedia-as-an-ontology).

Based on the performance aware architecture of MediaWiki it does not support any kind of user modelling based on the tracking of click streams or search interests. Nevertheless MediaWiki does provide a user management feature including a user profile base on the login name provided in order to edit or create MediaWiki entries.

Finally the MediaWiki extension *SocialProfile* allows the adding of Avatars, User Boards, extended user profile information and different types of user levels. This extension does extend the initial MediaWiki user profile although it does not provide any user modelling based on user behaviour.

Policies and Workflows

Related to the vast amount of users collaborating in different MediaWiki based Wikis it is essential that clear guidelines and policies related to user behaviour are followed. These guidelines mainly related to the revision and creation of articles. The most famous policies and guidelines are Wikipedias *Five Pillars* describing the top five policies that have to be obeyed when using the MediaWiki implementation Wikipedia.[138]

Localisation

Throughout the WikiMedia community a strong emphasis was put on multilingual aspects. Although Wikipedia as main WikiMedia implementation is available in over 100 languages, it has to be noted that each Wikipedia is unique in each language. It therefore does not provide any one to one translations of articles.[139]

Web Service and third part API integration

MediWiki provides a rich API for third party usage of the MediaWiki databases. The API enables the querying and retrieval of MediaWiki repositories.

References of Study

Drupal. (2009, February 25). *Drupal.org*. Retrieved February 25, 2009, from http://drupal.org/getting-started/before/overview

---

[138] Please note http://en.wikipedia.org/wiki/Wikipedia:Five_pillars or http://en.wikipedia.org/wiki/Wikipedia:Policies_and_guidelines for detailed description

[139] Please note http://www.mediawiki.org/wiki/Localisation_statistics for up to date statistics on localised MediWiki wikis

http://en.wikipedia.org/wiki/Wikipedia. (n.d.). *http://en.wikipedia.org/wiki/Wikipedia*. Retrieved February 25, 2009, from http://en.wikipedia.org/wiki/Wikipedia.

http://semanticsearch.org/. (n.d.). Retrieved February 25, 2009, from http://semanticsearch.org/: http://semanticsearch.org/

vanDyk, J. K. (2008). Pro Drupal Development. (p. 2). Apress Publishing.

Vidgen, S. G. (2002). Content, content, everywhere... time to stop and think? The process of web content management. (pp. Vol. 13, No. 2). Computer and Control Engineering Journal.

Volkel, M. K. (2006). Semantic Wikipedia. *WWW '06: Proceedings of the 15th international conference on World Wide Web* (pp. pp. 585-594). New York, NY, USA: ACM Press.

Appendix II – Online Survey to investigate the browsing experience of users in an across different Web-based Content Management Systems

Section I: General questions about browser usage

Question I.a:

Please indicate your browser usage?

Answer Options	Always	Very often	Frequently	Sometimes	Occasionally	Rarely	Never	Response Count
Firefox	14	10	2	1	0	1	0	28
Internet Explorer	3	2	1	2	5	9	3	25
Chrome	2	4	3	0	0	3	10	22
Safari	0	2	2	0	3	6	9	22
Opera	0	1	0	1	1	0	14	17
Flock	0	0	1	0	0	0	16	17
Other browsers not mentioned								3
							answered question	29
							skipped question	1

Question I.b:

Do you use different browsers for different goals/tasks/interests? (e.g. Flock for social sites and Firefox for everything else)

Answer Options	Response Percent	Response Count
Yes	44.8%	13
No	55.2%	16
If yes, please briefly state the browser(s) related to the type of task		13
answered question		29
skipped question		1

Question I.c:

Do you save your browsing session when exiting your browser (if available)?

Answer Options	Response Percent	Response Count
Yes	60.7%	17
No	39.3%	11
If it depends on the goal/task please briefly specify your usage		9
answered question		28
skipped question		2

Section II: General Questions about Web-based Content Management System usage

Question II.a:

Please specify your usage of WCMS and WCMS implementations					
Answer Options	**More than 5 times per day**	**1-5 times per day**	**Once in a while**	**Never**	**Response Count**
Seeking general information (e.g. History, Travel	6	10	6	1	23
Seeking specific information (e.g. technical	5	14	4	0	23
Seeking flimsy information (e.g. just surfing around)	3	8	10	2	23
Adding/editing a specific entry (e.g. blogs, articles etc.)	2	7	10	4	23
Adding/editing meeting minutes	0	3	12	8	23
Creating web pages/sites	1	6	10	6	23
Development of templates	0	3	8	12	23
Development of modules	0	1	10	12	23
API usage	1	1	10	11	23
Other usgage not listed					2
				answered question	**23**
				skipped question	**7**

Question II.b:

Please specify how you access the information of WCMS implementations							
Answer Options	**Very often**	**Frequently**	**Sometimes**	**Occasionally**	**Rarely**	**Never**	**Response Count**
Via search engine	12	8	2	1	0	0	23
Via browser based bookmarks	2	2	9	2	3	5	23
Via the search field within a browser	5	2	6	5	2	3	23
Via the search field within a WCMS	5	7	2	5	3	1	23
Via browser plug-ins (e.g. toolbars, delicious etc.)	3	2	2	3	8	5	23
Via links from other pages	1	8	5	4	4	1	23
By typing the URL into the address bar	3	3	6	3	3	5	23
If it depends on the task/goal or the WCMS implementation, please specify							2
						answered question	**23**
						skipped question	**7**

Question II.c:

When using a WCMS implementation (e.g. Wikipedia) are you generally aware of the underlying WCMS (e.g. MediaWiki)?		
Answer Options	**Response Percent**	**Response Count**
Yes	73.9%	17
No	26.1%	6
	answered question	**23**
	skipped question	**7**

Question II.d:

Do you use Browser plug-ins that have an effect on the usage of specific WCMS implementations (E.g. SmarterWiki, Passwordmanagers, dictionaries		
Answer Options	**Response Percent**	**Response Count**
Yes	4.3%	1
No	95.7%	22
If yes, please state the name(s) of the plug-in(s)		0
	answered question	**23**
	skipped question	**7**

Question II.e:

Please specify your navigation pattern when following an interest/goal/intent in one specific WCMS implementation, but over several pages.								
Answer Options	**Always**	**Very often**	**Frequently**	**Sometimes**	**Occasionally**	**Rarely**	**Never**	**Response Count**
I follow a small number of links on a page.	1	10	4	8	0	0	0	23
I tend to not go back to pages I have already visited.	0	1	7	7	4	4	0	23
I follow a large number of links on a page.	0	2	4	8	6	3	0	23
I tend to jump back to the pages I have already visited.	0	4	7	7	3	2	0	23
If I think a link could be relevant, I open it in a different	5	10	2	4	2	0	0	23
If I think a link could be relevant, I check immediately if	1	4	6	9	2	1	0	23
If it depends on the interets/goal/intent, please provide a brief example								2
							answered question	**23**
							skipped question	**7**

Question II.f:

How do you organise your browsing experience when following an interest/goal/intent in one specific WCMS implementation, but over several pages (e.g. the history of Dublin and its sights)?								
Answer Options	**Always**	**Very often**	**Frequently**	**Sometimes**	**Occasionally**	**Rarely**	**Never**	**Response Count**
Tabs	13	8	1	1	0	0	0	23
Windows	1	1	2	3	3	8	4	22
Different Browsers	1	0	0	1	3	5	12	22
Back/Forward Button (i.e. no new tab/window/browser	1	0	4	4	6	5	3	23
If it depends on the interets/goal/intent, please provide a brief example								1
							answered question	**23**
							skipped question	**7**

Question II.g:

Please specify your content type preference in WCMS implementations?							
Answer Options	**Very often**	**Frequently**	**Sometimes**	**Occasionally**	**Rarely**	**Never**	**Response Count**
If I have a choice I prefer Images	3	7	11	2	0	0	23
If I have a choice I prefer Text	11	9	3	0	0	0	23
If I have a choice I prefer Video	0	1	9	3	8	2	23
If I have a choice I prefer Speech	0	1	3	2	11	6	23
If it depends on the WCMS implementation, please specify							1
						answered question	**23**
						skipped question	**7**

Question II.h:

Do you experience any of the following hypertext/web phenomena when using a WCMS implementation?							
Answer Options	**Very often**	**Frequently**	**Sometimes**	**Occasionally**	**Rarely**	**Never**	**Response Count**
Lost in hyperspace (e.g. not sure which links I have	1	1	10	6	5	0	23
Cognitive overload (e.g. Too many choices/Too much	2	3	10	4	4	0	23
Development of a mental model of navigation paths	4	5	6	3	3	1	22
						answered question	**23**
						skipped question	**7**

Question II.i:

Please specify your usage of MediWikis side bar navigation aid which is usually grouped by topic.							
Answer Options	**Very often**	**Frequently**	**Sometimes**	**Occasionally**	**Rarely**	**Never**	**Response Count**
Navigation	3	4	4	3	6	2	22
Interaction	1	2	5	1	10	4	23
Toolbox	1	0	6	3	8	5	23
Languages	1	2	1	3	6	10	23
Other MediWiki Toolboxes not mentioned (e.g. KDEG Links)							0
						answered question	**23**
						skipped question	**7**

Question II.j:

Does the underlying WCMS have an influence on your query (e.g. I mostly use general keywords in a MediaWiki based WCMS, but very specific keywords in a		
Answer Options	**Response Percent**	**Response Count**
Yes	26.1%	6
No	73.9%	17
If it depends on the WCMS implementation or topic, please specify		3
answered question		**23**
skipped question		**7**

Section III: Questions related to the usage of different WCMS for the same interest/goal/intent

Question III.a:

How often do you use different WCMS implementations in seeking the same interest/goal/intent (e.g. Wikipedia and WikiHow)?		
Answer Options	**Response Percent**	**Response Count**
Very often	13.0%	3
Frequently	26.1%	6
Sometimes	21.7%	5
Occasionally	8.7%	2
Rarely	21.7%	5
Never	8.7%	2
answered question		**23**
skipped question		**7**

Question III.b:

If possible, please briefly specify an interest/goal/intent for which you use or did use different WCMS implementations?	
Answer Options	**Response Count**
	9
answered question	**9**
skipped question	**21**

Question III.c:

How do you organise your browsing experience when following an interest/goal/intent across different WCMS implementation (e.g. reading about Dublin in Wikipedia and then reading a blog in a Wordpress based WCMS about Dublin)?								
Answer Options	**Always**	**Very often**	**Frequently**	**Sometimes**	**Occasionally**	**Rarely**	**Never**	**Response Count**
Tabs	12	6	2	1	1	0	1	23
Windows	1	2	3	1	3	8	4	22
Different Browsers	1	0	0	0	3	6	12	22
Back/Forward Button (i.e. no new tab/window/browser	1	0	1	6	2	5	6	21
If it depends on the goal, please provide a brief example								1
							answered question	**23**
							skipped question	**7**

Question III.d:

Please specify your navigation pattern when following an interest/goal/intent over different WCMS implementations?								
Answer Options	**Always**	**Very often**	**Frequently**	**Sometimes**	**Occasionally**	**Rarely**	**Never**	**Response Count**
I don't mind if a link is pointing out of the WCMS	5	5	6	3	2	2	0	23
If I think a link pointing to a different WCMS	7	9	1	3	2	1	0	23
If I think a link pointing to a different WCMS	1	3	2	11	3	2	0	22
If it depends on the interets/goal/intent, please provide a brief example								1
							answered question	23
							skipped question	7

Question III.e:

Do you experience any of the following hypertext/web phenomena when using different WCMS implementation for the same interest/goal/intent?							
Answer Options	**Very often**	**Frequently**	**Sometimes**	**Occasionally**	**Rarely**	**Never**	**Response Count**
Lost in hyperspace (e.g. not sure which WCMS I	2	0	10	5	6	0	23
Cognitive overload (e.g. Too many choices/Too much	1	3	7	7	5	0	23
Development of a mental model of navigation paths	3	5	3	5	4	2	22
						answered question	23
						skipped question	7

Section IV: Questions about personalisation in WCMS

Question IV.a:

Please indicate the usefulness of following personalisation categories in and across WCMS?					
Answer Options	**Very Useful**	**Useful**	**Possibly useful**	**Not useful**	**Response Count**
Profile/User settings based personalisation set by the	6	9	7	0	22
Behavior based personalisation set by the WCMS (e.g.	4	13	4	1	22
Collaboration based personalisation set by a community	7	9	5	1	22
Other peronalisation categories not mentioned					1
				answered question	23
				skipped question	7

Question IV.b:

Please indicate the usefulness of following personalisation in and across WCMS					
Answer Options	**Very Useful**	**Useful**	**Possibly useful**	**Not useful**	**Response Count**
Assisting the usage of different WCMS (e.g. information	8	5	9	0	22
Link highlighting based on relevance/previous	4	13	6	0	23
Content highlighting based on relevance/previous	5	13	5	0	23
Enable/Disable functions that are rarely used	2	4	13	4	23
Enable/Disable navigation aids (mostly in the left side	0	8	9	6	23
Assisting the adding/editing of content	2	14	7	0	23
Other personalised applications you think could be useful, but are not mentioned					3
				answered question	23
				skipped question	7

Question IV.c:

Would you allow a WCMS to store and use your user profile (i.e. you have generated it yourself)?		
Answer Options	**Response Percent**	**Response Count**
Yes	86.4%	19
No	13.6%	3
If not sure, please specify		4
	answered question	**22**
	skipped question	**8**

Question IV.d:

Would you allow a WCMS to create a user model based on you browsing behavior (i.e. the WCMS generates it)?		
Answer Options	**Response Percent**	**Response Count**
Yes	90.9%	20
No	9.1%	2
If not sure, please specify		4
	answered question	**22**
	skipped question	**8**

Question IV.e:

Would you allow a third party service to access your profile/user model in a specific WCMS?		
Answer Options	**Response Percent**	**Response Count**
Yes	43.5%	10
No	56.5%	13
If not sure, please specify		4
	answered question	**23**
	skipped question	**7**

Question IV.f:

Would you allow a third party service to use your profile/user model across different WCMS?		
Answer Options	**Response Percent**	**Response Count**
Yes	47.8%	11
No	52.2%	12
If not sure, please specify		3
	answered question	**23**
	skipped question	**7**

Question IV.g:

If not yet stated, please provide a brief example in which personalisation within or across different WCMS would be of	
Answer Options	**Response Count**
	8
answered question	**8**
skipped question	**22**

Question IV.h:

If you approve specific follow up questions to this survey, please state your name or email address.	
Answer Options	**Response Count**
	11
answered question	**11**
skipped question	**19**

Appendix III – First Experiment (Closed Domains)

A.1. Tasks

Please note that all tasks are voluntary. You may choose to skip questions without penalty and without providing a reason.

Task 1: Backup

Norton 360 contains capabilities to back up your PC.

You would like to:

1) Gather basic introductory information about this feature

and **2)** learn about how to create such backups.

3) You have an old backup called 'Henry' that you want to delete and make a completely new backup. How do you do this?

4) During this process, you encounter some problems, and you would like to find some information on this particular problem.

Task 2: Updating

You get the following error: "Module:8920, Error:201, ANTIVIRUS DEFINITIONS NOT UPDATING".

1) Which feature is responsible for updating definitions (please provide an introduction to this feature)?

2) For which kinds of threats do you need definition updates?

3) Can you find some solutions to the error message?

Task 3: Pulse updates vs LiveUpdate

1) What is the difference between LiveUpdate and Pulse updates?

2) Can you find some solutions to the following error: "Pulse Updates from Symantec Security failed to complete"?

Task 4: Silent mode

1) What happens when Silent Mode is turned enabled?

a. The computer's sound is muted for a specified duration.

b. The computer goes to Standby mode for a specific duration.

c. The product suppresses alerts and certain background activities for a specific duration of time.

d. All features of the product are disabled for a specific duration of time.

2) In which of the following situations is Automatic Silent Mode enabled? (Check all that apply)

a. When the user launches a full screen application such as a full screen game.

b. In the event of a CD or a DVD burn.

c. When you open more than 5 applications.

d. When the screen saver comes up.

3) You can disable Automatic Silent Mode from the Settings user interface.

a. True

b. False

Task 5: Identity Protection

1) Which of the following activities does Identity Protection perform?

a. Safeguards against online identity theft

b. Scans the incoming and the outgoing emails for malicious content

c. Blocks the fake Web sites and Crimeware

d. Stores and encrypts your passwords to prevent accidental disclosure to unknown sites or unauthorized sites

2) Which of the following features does Norton Safe Web provide?

a. Offers a Web site rating service that extends protection to everyday online activities: search, browse, transact, and interact

b. Provides a trusted visual indicator on search results pages of search engines

c. Protects from the sites that can infect the computers and misuse user's personal information

d. Protects the host file from any changes

3) Which of the following statements are true about a Portable profile?

a. You can use the same portable Identity Safe profile with different user accounts on the same computer as well as on different computers where a compatible Identity Safe component exists in a Norton product

b. Any device that appears as a "removable disk" in Windows, with at least 5MB free space, and has a drive letter can be supported

c. The support includes USB flash drives, portable USB hard drives, cameras, MP3 players, cell phones, Zip drives

d. Optical drives and a fixed hard drive are also supported if user has rights to write on them

Task 6: Scheduling Tasks

1) What is the difference between configuring idle time automatic tasks and scheduling tasks?

2) Can you schedule automatic Backups?

3) How do you specify the idle time duration for automatic tasks?

4) You wish to disable idle time scans. Where do you do this?

Task 7: Miscellaneous Problems

Please try to solve the following problems:

1) "How do I get my norton toolbar back?"

2) "I am having problems with live update: unable to connect to network because proxy server"

3) "How to restore files from dvd?"

A.2. Task performance metrics (through logging)

User Action Logging

1. Task completion time
2. Number of queries per task
3. Number of clicks to answer question
4. Number of pages viewed

User Information Foraging

5. Amount of collected information
6. Diversity of collected information (forums/documentation)

User Sense Making

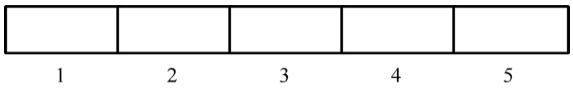
7. Amount of information selected for “final answer”
8. Task performance (answer accuracy & completeness)
9. Ratio of successes to failures

A.3. Pre-test Questionnaire

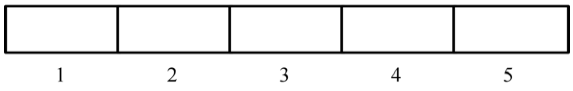
Please note that all questions are voluntary. You may choose to skip questions without penalty and without providing a reason.

Strongly disagree — Strongly agree

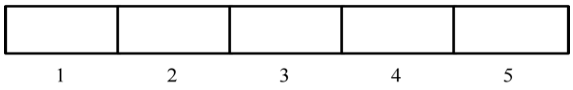
1. I am aware of the product Norton 360 and its features.



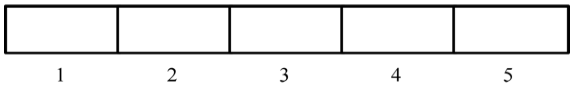
2. I use Norton 360 frequently.



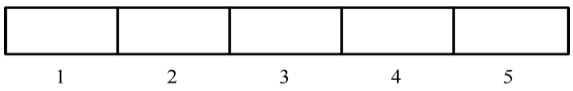
3. I am an expert on Norton 360.



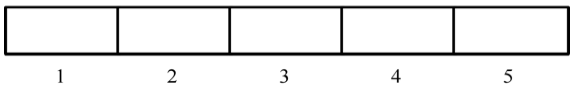
4. I often consult product manuals to find information about software features in general.



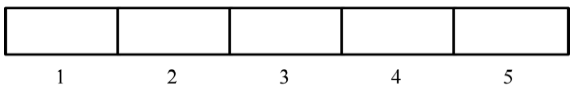
5. I often consult community forums to find information about software features in general.



6. I often use general web search engines to find information about software features in general.



7. I often consult product manuals to find problem solutions.



8. I often consult community forums to find problem solutions.

9. I often use general web search engines to find problem solutions.

10. I have used adaptive systems in the past.

A.4. Usability Questionnaires

Please note that all questions are voluntary. You may choose to skip questions without penalty and without providing a reason.

System Usability Scale (SUS)

	Strongly disagree				Strongly agree
1. I think that I would like to use this system frequently	1	2	3	4	5
2. I found the system unnecessarily complex	1	2	3	4	5
3. I thought the system was easy to use	1	2	3	4	5
4. I think that I would need the support of a technical person to be able to use this system	1	2	3	4	5
5. I found the various functions in this system were well integrated	1	2	3	4	5
6. I thought there was too much inconsistency in this system	1	2	3	4	5
7. I would imagine that most people would learn to use this system very quickly	1	2	3	4	5

8. I found the system very cumbersome to use

9. I felt very confident using the system

10. I needed to learn a lot of things before I could get going with this system

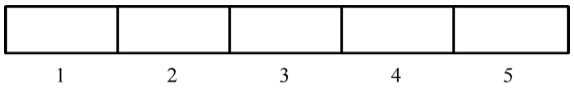
Please note that all questions are voluntary. You may choose to skip questions without penalty and without providing a reason.

User Satisfaction regarding system-specific features

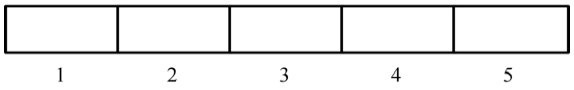
Strongly Strongly

disagree agree

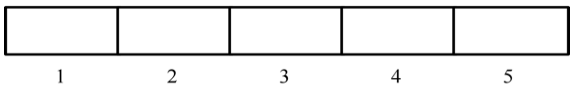
1. I found the adaptive hints were relevant related to my current browsing behaviour.



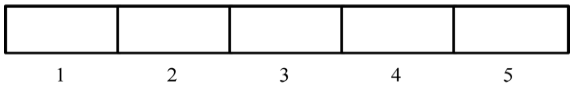
2. I found the adaptive hints were irrelevant.



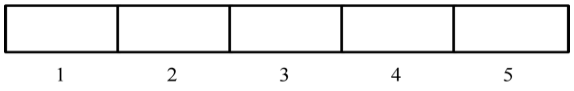
3. I found the presentation of the adaptive hints was helpful.



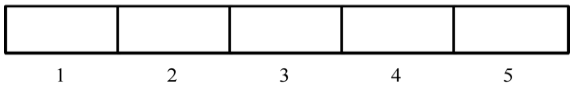
4. I found the presentation of the adaptive hints was too intrusive.



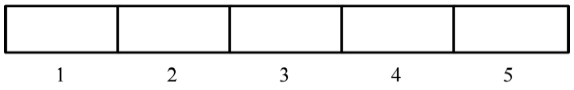
5. I found the presentation of the adaptive hints distracting.



6. The content indicated by the adaptive hints was matching my expectations.



7. There were too few adaptive hints.



8. The adaptive hints did not help me solve the tasks.

9. It was helpful to know which terms are related to the adaptive hint.

10. The indicated terms were not related to the underlying content

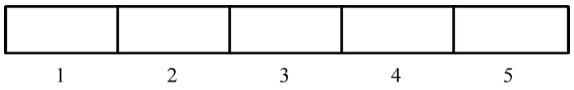
Please note that all questions are voluntary. You may choose to skip questions without penalty and without providing a reason.

General User Satisfaction, Reaction & Comments

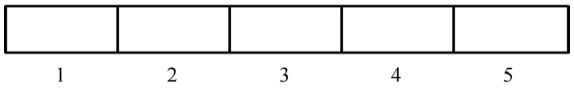
Strongly Strongly

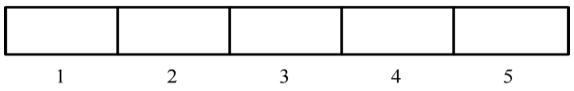
disagree agree

1. I had to use the Drupal search function a lot before I found interesting content.

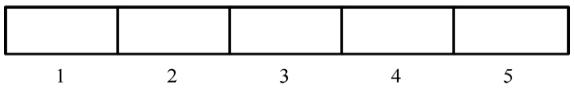


2. I did well on the different tasks.

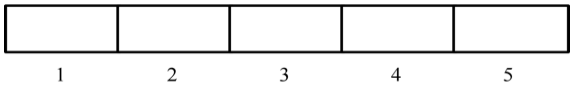


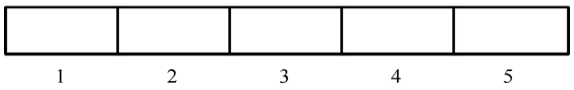


3. Overall, I am satisfied with the system performance, assistance and guidance.

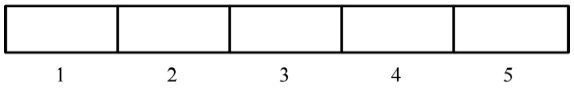


4. The system guided me towards more personally relevant content.

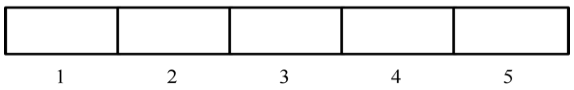


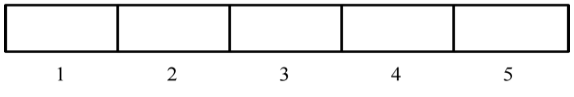


5. I found that the interaction with the system motivated me to spend more time browsing.

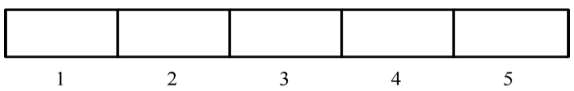


6. I found the interaction with the system helpful in understanding my personal browsing behaviour.





7. I would like to use the system in a more open browsing scenario.



8. I would like to know more about my browsing behaviour by viewing my user model.

9. I would like to be able to control and change my user model to influence the adaptive behaviour.

10. I would like to be able to manipulate the adaptive rules allowing more effective adaptive hints based on my individual overall browsing interests.

11. What features/characteristics did you like most about the system?

12. What features/characteristics did you like least about the system?

13. What kind of adaptivity other than link annotation would you consider useful?

14. What would you consider the main challenges in using this system on the open web?

15. Would you have privacy concerns using this system on the open web?

A.5 Additional Questions

1. Was the approach useful in assisting tasks?

2. Would users like to use the approach in a more open web-browsing scenario?

3. Was adaptive link presentation was perceived as too intrusive?

4. Was the adaptive link presentation perceived as distracting?

5. Was the user constantly aware of the personalised feature?

6. Did the user think not enough recommendations were provided?

7. Did the approach reduce the feeling of disorientation?

8. Would users like to add more information to the model such as age, language preferences etc.

9. Would users like more control over the approach?

10. Were the users satisfied with the performance of the personalised system pair in comparison to the non-personalised?

Selected data outputs

Overall Average Time per Task

			All in			without outlier			
Overall Average Time System A (non adaptive)			1:16:25			1:23:53			
Overall Average Time System B (adaptive)			1:29:30			1:40:03			
delta			0:13:05			0:16:10			
	All in	average per user	Max_Value	Min_Value		without outlier	average per user	Max_Value	Min_Value
Overall Time Task 1 System A	3:07:14	0:05:21	0:43:40			3:04:21	0:05:16		
Average Time per task	0:20:48	0:00:36				0:26:20	0:00:45		
Overall Time Task 1 System B	3:31:33	0:06:03				3:31:33	0:06:03		
Average Time per task	0:30:13	0:00:52				0:30:13	0:00:52		
Overall Time Task 2 System A	1:49:20	0:03:07				1:49:16	0:03:07		
Average Time per task	0:13:40	0:00:23				0:15:37	0:00:27		
Overall Time Task 2 System B	2:38:06	0:04:31				2:33:34	0:04:23		
Average Time per task	0:17:34	0:00:30				0:25:36	0:00:44		
Overall Time Task 3 System A	3:02:33	0:05:13				3:02:33	0:05:13		
Average Time per task	0:22:49	0:00:39				0:22:49	0:00:39		
Overall Time Task 3 System B	3:58:32	0:06:49				3:57:28	0:06:47		
Average Time per task	0:23:51	0:00:41				0:26:23	0:00:45		
Overall Time Task 4 System A	2:52:05	0:04:55				2:52:05	0:04:55		
Average Time per task	0:19:07	0:00:33				0:19:07	0:00:33		
Overall Time Task 4 System B	2:22:49	0:04:05				2:22:49	0:04:05		
Average Time per task	0:17:51	0:00:31				0:17:51	0:00:31		

Overall content views per task

Overall Content Loaded per Task	Task 1	Task 2	Task 3	Task 4	Total	Average	Term Model Views	
System 1	80	61	62	91	294	73.5		
System 2	38	17	19	17	91	22.75		
System 3	121	68	103	63	355	88.75	59	
System 4	32	15	30	32	109	27.25	86	
Total	271	161	214	203	849			
Average System 1 and 2	59	39	40.5	54		48.125		
Average System 3 and 4	76.5	41.5	66.5	47.5		58		
Average Total								

	Personalised System Pair	Non Personlised System Pair		Personalised Manual	Non Personalised Manual		Personalised Forum	Non Personalised Forum
Average	116	96.25		88.75	73.5		27.25	22.75
Task 1	153	118		121	80		32	38
Task 2	83	78		68	61		15	17
Task 3	133	81		103	62		30	19
Task 4	95	108		63	91		32	17

Appendix IV – Two-Phased Design Study: Relevance, Usability and Visual Assistance

Phase 1 Questionnaires: Investigation of Visual Assistance Types

Mock-up Evaluation Reaction Cards

Date:
User ID:
Visualisation Type:
Please note that all questions are voluntary. You may choose to skip questions without penalty and without providing a reason.
REACTION CARDS Please choose **5** terms from this list that most closely reflect your initial reaction to the composition.

Annoying	Creative	Friendly	Organized	Straight Forward
Appealing	Desirable	Frustrating	Overbearing	Stressful
Attractive	Difficult	Gets in the way	Overwhelming	Too Technical
Boring	Distracting	Impersonal	Patronizing	Trustworthy
Business-like	Dull	Impressive	Poor quality	Unapproachable

Busy	Empowering	Inconsistent	Powerful	Unattractive
Calm	Energetic	Inspiring	Predictable	Unconventional
Clean	Engaging	Intimidating	Professional	Understandable
Clear	Enthusiastic	Intuitive	Relevant	Undesirable
Comfortable	Exciting	Inviting	Rigid	Unpredictable
Compelling	Expected	Meaningful	Satisfying	Unrefined
Complex	Familiar	Motivating	Simplistic	Usable
Comprehensive	Fragile	Not Valuable	Sterile	Useful
Confusing	Fresh	Ordinary	Stimulating	

Explanations:

Mock up Evaluation – Questions to Visual Assistance

Date:
User ID:
Visualisation Type:
Please note that all questions are voluntary. You may choose to skip questions without penalty and without providing a reason.

1. I would follow link annotations that are based on my cross-site browsing experience, frequently.	
☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree	
2. I would like the system to influence the overall experience of the website more directly, such as displaying relevant images and pushing relevant content to the front page.	
☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree	
3. I would like the system to highlight terms that are relevant in relation to my cross-site browsing experience.	
☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree	
4. Can you think of any other cross-site support mechanisms that you would perceive as useful as a user? If yes, please explain which ones and why they would be useful.	

5. I would like the system to recommend links that are related to my cross-site interest, on the side of the webpage.	
☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree	

Mock up Evaluation – Questions to Overall Approach

Date:	
User ID:	
Visualisation Type:	
Please note that all questions are voluntary. You may choose to skip questions without penalty and without providing a reason.	
Question	Comment
1. I would use this service frequently. If agree, why and if disagree, what would need to be addressed?	
☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree	
2. Describe use cases in which you think this approach will be most useful.	

3. Are there any concerns related to the overall approach? If yes, please explain the main concern.	
4. How would you measure/evaluate the usefulness of this approach?	

Phase 1 Findings: Investigation of Visual Assistance Types

In the following the three different topics link annotation, authentication and term visualisation is discussed.

Link Visualisation

To assess the link visualisation of the user four different visualisation strategies were assessed: (1) Link Highlighting; (2) Link annotation with a star symbol; (3) Link annotation with a circular symbol and (4) a small dotted bar.

For each strategy different levels of link relevance were assessed. In link highlighting the relevance was assesses based on colour intensity. In link annotation with star symbols the colour scheme bronze, silver and gold was chosen to indicate relevancy. For link annotation with a circular symbol Harvey balls[140] were chosen. Finally, for the dotted bar indicates relevancy based on filling the bars three separate sections. Each user was presented with at least three different website examples and colours to avoid a bias towards a website specific colour scheme. The following discussion is a summary of the user's feedback.

Link Highlighting

The overall reaction of the users was positive. Users commented that it would work well if only few elements are highlighted. However, the relevancy was not obvious enough. For this users indicated that a numbering may be useful.

The reaction card feedback was mostly positive with 32 positive words and 10 negative words indicated.

Reaction Cards (up to 5 terms per participant):

Used: Appealing (2), Attractive (1), Boring (1), Compatible (1), Consistent (2), Convenient (1), Clear (1), Confusing (2), Disruptive (1), Distracting(1), Dull (1), Efficient (2), Effective (5), Familiar (2), Gets in the

[140] http://en.wikipedia.org/wiki/Harvey_Balls

way (1), Helpful (2), Intuitive (3), Integrated(3), Inviting (1), Noticeable (3), Not Valuable (1), Ordinary (1), Organized (1), Overwhelming (1), Unattractive (3), Useful (1), Valuable (1)

Asked if the visual assistance type is visible two 2 participants disagreed, 2 neutral, 3 agreed and 2 strongly agreed. Over 50% of the users agreed that the link highlighting is intuitive. Asked if link highlighting is confusing user reacted balanced with 2 users strongly disagreeing, 3 users disagree, but 2 users agreeing and 2 user disagreeing. Most users disagreed that link highlighting is too intrusive. More than 50% of the users agreed that link highlighting would influence their browsing behaviour in a positive way and that the strategy was not distracting.

User liked most that the strategy is simple, clean and that it may save time. Users disliked that it was not obvious and sometimes hard to see. Furthermore that the website may tend to overload id to many items were highlighted. As improvement users indicated a colour scheme that the user can select.

Link Annotation with a start symbol

Interestingly the star symbol resulted in the selection of 37 positive and 2 negative word reactions. Even though the initial reaction was positive most users indicated that the colour scheme was not clear and not intuitive.

Most users agreed that the visual assistance was noticeable, intuitive and that it was not perceived as confusing. Furthermore almost all users agreed that the strategy was not too intrusive, that it would affect browsing in a positive way and that it was not distracting.

User liked most that the star icon was not intrusive and not distracting. However, some users indicated that it did not work well over text and would work better on image links. Also the colour scheme was too subtle with bronze and gold both looking too close to yellow. Similar to link highlighting user would like to be able to choose the icon colour.

Link annotation with a circular symbol

Surprisingly the reaction of the users for the circular symbol was overwhelmingly positive with 46 positive word selections and 0 negative word selections. Most comments were made in relation to how

the circle is filled. The filling used reminded some users of a clock which indicated time and not relevancy.

More than 50% of the users strongly agreed that the visual assistance was visible and intuitive. More than 50% of the users strongly disagreed that the circle was confusing, too intrusive or distracting. All users agreed or strongly agreed that the strategy would positively affect their browsing.

User liked most that the colour was noticeable and that it contains a lot of information in a small space. Surprisingly some users mentioned that it felt personal. The only negative comment stated was that it was confusing, but this comment was only stated by one participant.

To improve the strategy participants indicated that the circle should fill horizontally depending on the level of relevancy.

Link annotation with bar (dots)

This strategy was perceived as most negative with only 5 positive words selected and 19 negative.

Two users strongly disagreed that the visual assistance was noticeable; 2 answered neutral, 1 agreed and 1 strongly agreed. Similar mixed the answers related to how intuitive the strategy is (1 strongly disagree, 1 disagree, 2 neutral, 1 agree and 2 strongly agree). Asked if the strategy is confusing 2 agreed, 2 neutral, 1 agreed and 2 strongly agreed. In relation to intrusiveness most users either agreed or strongly agreed that the strategy was non-intrusive (1 neutral answer). In relation to the strategy influencing the user's browsing in a positive manner 1 participant strongly agreed, 2 agreed, 3 disagreed and 1 remained neutral. Finally, over 50% of the participants agreed that the strategy was not distracting.

Participants like that it was clear and simple. However, most comments were negative, such as that the icon was too small and it was hard to identify the relevancy.

Visual assistance *feedback on selected link visualisation*

Based on the positive feedback and reaction the link annotation circle was selected for the final experiment. Furthermore, the relevancy filling was change from quarters to horizontal filling to void the feeling of time

After presenting the final strategy users reacted mostly positive with selecting 57 positive words and only 2 negative words.

Asked if the visual assistance is noticeable and intuitive all users either agreed or strongly agreed. All participants, bar one neutral answer, disagreed that the strategy is confusing. All users, bar one, strongly disagreed that the strategy is too intrusive. Finally, all users either disagreed or strongly disagreed that the strategy was distracting.

In relation to the general feedback of the users five concluding questions were asked:

First, the participants were asked if they would like to follow the selected annotation if it is based on the Cross Site browsing experience. Two participants remained neutral and commented that they would need to see a live use case. Four participants agreed and three strongly agreed. Some users expressed the need to allow disabling if necessary. The second question related to the level of invasiveness the CSP approach should have. For example, should the approach be able to create a more personal experience by loading a specific page and not only highlight the link. Surprisingly over 50% of the users agreed to a more invasive integration of the approach beyond simple link annotation. Asked if the participants would like keywords to be highlighted in the recommended text also over 50% of the users agreed with some mentioning that they would like to be able to enable and disabled the displaying.

Finally, the participants were asked to provide feedback related to the overall approach of CSP. Four questions were asked. The first question asked if the participant would use a CSP approach frequently. Surprisingly most participants agreed (6 strongly agreed, 3 agreed and one neutral. Asked what in what use cases CSP would be most usefulness sports, news, academia and tourism was mentioned. Asked about concerns most answered related to data privacy.

Phase 2 Questionnaire: Behavioural Tracking and Recommendation Creation

A.1. Tasks

Please note that all tasks are voluntary. You may choose to skip questions without penalty and without providing a reason.

Task 1: Cross-Site Browsing

Familiarise yourself with a pre-selected interest by browsing across the provided website. The pre-selected interest is:

a) Health, Fitness through Golfing

	Comment
I have a genuine interest in the selected topic.	
☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree	
	Comment
I am familiar with the pre-selected topic.	
☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree	

Provide up to 5 terms that describe the content browsed.	1. 2. 3. 4. 5.
Indicate the relevancy of each term between 0-10 with 10 being the most relevant.	1. [0-10]: 2. [0-10]: 3. [0-10]: 4. [0-10]:

	5. [0-10]:

Browse to your interest model and discuss it based on following statements:

	Comment
The terms indicated by the service are related to my current browsing interest.	
☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree	
The size (indication of relevancy) of the terms correspond to my current browsing interest.	
☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree	
Indicate the three most relevant terms.	1. 2.

	3.
Indicate the three least relevant terms.	1. 2. 3.

Browse to the following location website: Killarney.

Rate each link in the recreation section between 0 and 10 with 10 being most relevant.	
Castlerosse Golf Course	[0-10]:
Deerpark Golf and Pitch & Putt Club	[0-10]:

Gap of Dunloe Cycle Route	[0-10]:
Gap of Dunloe Walking Tours	[0-10]:
Humphry Fishing Lakes Killarney	[0-10]:
Killarney Cycling Trail	[0-10]:
Killarney Golf And Fishing Club	[0-10]:
Killarney Guided Walks	[0-10]:
Killarney Historic Town Walk	[0-10]:
Killarney Walking Trail	[0-10]:
Mangerton Mountain Hill Walk	[0-10]:
Muckross And Dinis Walking Tour	[0-10]:
Ross Golf Club	[0-10]:
The Kerry Way Walk	[0-10]:

Browse to the following location website: Dingle.

Rate each link in the recreation section between 0 and 10 with 10 being most relevant.	
Charter Fishing Boat Molly'O	[0-10]:
Dingle Bay Charters Sea Angling	[0-10]:
Dingle Pitch & Putt	[0-10]:
Láthair Spóirt an Daingin Golfing	[0-10]:
Paddy's Rent A Bike	[0-10]:
The Dingle Way Walks	[0-10]:
The Kerry Mountains Hill Walks and Tours	[0-10]:
The Mysteries of Dingle Cycle Route	[0-10]:

Browse to the following location website: Youghal.

Rate each link in the recreation section between 0 and 10 with 10 being most relevant.	
Claycastle Pitch And Putt Club	[0-10]:
Guided Tours Of Olde Youghal	[0-10]:
Lough Aderra Fishery	[0-10]:
South Coast Charter Angling	[0-10]:
The Blackwater Valley Cycle Route	[0-10]:
The East Cork – Cork City to Youghal Cycle Route	[0-10]:
Youghal Deep Sea Angling	[0-10]:
Youghal Golf Club	[0-10]:

A.2. Task performance metrics (through logging)

User Action Logging

1. Task completion time
2. Number of queries per task
3. Number of clicks to answer question
4. Number of pages viewed
5. Time spent on page
6. Mouse actions on page
7. Keyboard actions on page

Phase 2 Findings: Behavioural Tracking and Recommendation Creation

To assess the background of the users the first question assessed if the participants have genuine interest in the topic and the second question asked if the participants had prior knowledge in the topic. The first question indicated a good balance with 1 participant strongly disagreeing, 3 users disagreeing, 4 neutral and 2 agreeing. The second question indicated that the participants have prior knowledge about golf with 2 participants strongly disagreeing, 4 disagreeing and 4 agreeing.

Next the participants were asked to familiarise themselves with the topic of golf by reading content on a website related to health and fitness benefits of golf. After indicating the terms the participants were asked to indicate relevancy between 0 and 10. In the following an excerpt of the terms with an average score attached.

Term	Mentions	Score
Golf	8	10/10/10/7/10/10/10/9
Fitness	7	8/10/9/10/10/10/7
Health	7	9/4/10/8/10/7/9
Sport	1	10
Old People	1	6
Calories	1	5
Walking	3	7/8/10
Research studies	2	7/10
Health	2	8/8

After rating the self-stated terms the participants were presented with the term cloud the service inferred based on their browsing behaviour. Asked if the resulting term cloud was relevant 2 participants disagreed, 3 neutral, 4 agreed and 3 strongly agreed. Asked if the term size that indicated the level of relevancy was accurate 2 participants disagreed, 3 neutral, 4 agreed and 2 strongly agreed. The participants were then asked to select the three most and least relevant terms. All 10 participants selected Golf, 4 selected fitness, 4 sports (other terms were among others golf equipment, human behaviour and recreation). Least relevant terms were Greg Norman with 5 participants, Business and Finance with 7 and machine press with 5 (other terms selected were among others mind body and

medical pharma). It should be noted that the term model was slightly different for each participant based on the different browsing behaviour. The initial feedback is promising though due to overlaps between the freely chosen terms of the users and the inferred terms of the service.

After this initial assessment the participants were asked to browse to three different location websites where local golf amenities were indicated in a link list (together with other locations related to cycling, walking and hiking).

For each website the participants were asked to judge the relevancy of the links showed at the side of the website. This user relevancy judgement was then compared with the calculated relevancy of the service. The following charts illustrate the outcomes. For each user the individual judgement is compared with the calculated score of the system. The difference is indicated in brackets (first the participants rating followed by the systems rating). The lower the difference the better the match. The golf related links are indicated in bold.

Link (Killarney)	User 1	User 2	User 3	User 4	User 5	User 6	User 7	User 8	User 9	User 10	Overall Sum
Castlerosse Golf Course	**5-4 (1)**	**5-4 (1)**	**9-4 (5)**	**5-4 (1)**	**9-4 (5)**	**10-4 (6)**	**9-4 (5)**	**8-4 (4)**	**6-4 (2)**	**10-4 (6)**	**36**
Deerpark Golf and Pitch & Putt Club	**5-1 (4)**	**3-1 (2)**	**9-1 (8)**	**4-1 (3)**	**8-1 (7)**	**2-1 (1)**	**9-1 (8)**	**6-1 (5)**	**8-1 (7)**	**2-1 (1)**	**46**
Gap of Dunloe Cycle Route	1-0 (1)	1-0 (1)	3-0 (3)	4-0 (4)	5-0 (5)	0-0 (0)	2-0 (2)	4-0 (4)	0-0 (0)	10-0 (10)	29
Gap of Dunloe Walking Tours	1-1 (0)	1-1 (0)	3-1 (2)	4-1 (3)	4-1 (3)	0-1 (1)	0-1 (1)	5-1 (4)	3-1 (2)	5-1 (4)	20

Humphry Fishing Lakes Killarney	0-0 (0)	1-0 (1)	2-0 (2)	2-0 (2)	4-0 (4)	0-0 (0)	2-0 (2)	3-0 (3)	0-0 (0)	0-0 (0)	14
Killarney Cycling Trail	1-1 (0)	1-1 (0)	3-1 (2)	2-1 (1)	6-1 (5)	0-1 (1)	6-1 (5)	3-1 (2)	2-1 (1)	0-1 (1)	18
Killarney Golf And Fishing Club	**7-10 (3)**	**6-10 (4)**	**8-10 (2)**	**7-10 (3)**	**8-10 (2)**	**9-10 (1)**	**9-10 (1)**	**9-10 (1)**	**9-10 (1)**	**8-10 (2)**	**17**
Killarney Guided Walks	1-1 (0)	1-1 (0)	3-1 (2)	3-1 (2)	6-1 (5)	0-1 (1)	7-1 (6)	5-1 (4)	4-1 (3)	2-1 (1)	24
Killarney Historic Town Walk	1-1 (0)	1-1 (0)	2-1 (1)	3-1 (2)	5-1 (4)	0-1 (1)	0-1 (1)	5-1 (4)	3-1 (2)	0-1 (1)	16
Killarney Walking Trail	1-1 (0)	1-1 (0)	3-1 (2)	2-1 (1)	4-1 (3)	0-1 (1)	5-1 (4)	5-1 (4)	2-1 (1)	0-1 (1)	17
Mangerton Mountain Hill Walk	1-1 (0)	1-1 (0)	3-1 (2)	2-1 (1)	6-1 (5)	0-1 (1)	4-1 (3)	5-1 (4)	1-1 (0)	0-1 (1)	17
Muckross And Dinis Walking Tour	1-0 (1)	1-0 (1)	3-0 (3)	2-0 (2)	7-0 (7)	0-0 (0)	7-0 (7)	5-0 (5)	1-0 (1)	2-0 (2)	29
Ross Golf Club	5-5 (0)	5-5 (0)	8-6 (2)	7-5 (2)	9-6 (3)	5-6 (1)	9-5 (4)	7-5 (2)	9-5 (4)	7-5 (2)	20
The Kerry	1-0	1-0	3-0	3-0	7-0	0 -0	7-0	5-0	2-0	3-0	30

Way Walk	(1)	(1)	(3)	(3)	(7)	(0)	(7)	(5)	(0)	(3)	

System and user rating for each link website 2 – indicate point distance

Link (Dingle)	User 1	User 2	User 3	User 4	User 5	User 6	User 7	User 8	User 9	User 10	Overall Sum
Charter Fishing Boat Molly'O	0-1 (1)	1-1 (0)	1-1 (0)	3-1 (2)	4-1 (3)	0-1 (1)	0-1 (1)	3-1 (2)	0-1 (1)	7-1 (6)	17
Dingle Bay Charters Sea Angling	0-1 (1)	1-1 (0)	1-1 (0)	2-1 (1)	4-1 (3)	0-1 (1)	1-1 (0)	3-1 (2)	0-1 (1)	7-1 (6)	15
Dingle Pitch & Putt	**5-10 (5)**	**5-10 (5)**	**9-10 (1)**	**7-10 (3)**	**8-10 (2)**	**5-10 (5)**	**9-10 (1)**	**5-10 (5)**	**8-10 (2)**	**4-10 (6)**	**35**
Láthair Spóirt an Daingin Golfing	**5-7 (2)**	**5-7 (2)**	**9-6 (3)**	**7-7 (0)**	**8-7 (1)**	**7-6 (1)**	**10-6 (4)**	**9-4 (5)**	**9-6 (3)**	**10-7 (3)**	**24**
Paddy's Rent A Bike	1-1 (0)	1-1 (0)	3-1 (2)	2-1 (1)	6-1 (5)	0-1 (1)	7-1 (6)	4-1 (3)	1-1 (0)	4-1 (3)	21
The Dingle Way Walks	1-1 (0)	1-1 (0)	3-1 (2)	2-1 (1)	7-1 (6)	0-1 (1)	6-1 (5)	6-1 (5)	2-1 (1)	5-1 (4)	25
The Kerry Mountains Hill Walks and Tours	1-1 (0)	1-1 (0)	3-1 (2)	2-1 (1)	6-1 (6)	0-1 (1)	7-1 (6)	6-1 (5)	0-1 (1)	6-1 (5)	27

The Mysteries of Dingle Cycle Route	1-1 (0)	1-1 (0)	2-1 (1)	2-1 (1)	7-1 (6)	0-1 (1)	7-1 (6)	4-1 (3)	0-1 (1)	5-1 (1)	20

System and user rating for each link website 3 – indicate point distance

Links (Youghal)	User 1	User 2	User 3	User 4	User 5	User 6	User 7	User 8	User 9	User 10	Overview Sum
Claycastle Pitch And Putt Club	**5-7 (2)**	**5-7 (2)**	**8-8 (0)**	**8-7 (1)**	**8-7 (1)**	**5-8 (3)**	**9-7 (2)**	**2-5 (3)**	**8-8 (0)**	**4-7 (3)**	**17**
Guided Tours Of Olde Youghal	1-1 (0)	1-1 (0)	2-1 (1)	2-1 (1)	6-1 (5)	0-1 (1)	7-1 (6)	5-1 (4)	2-1 (1)	6-1 (5)	24
Lough Aderra Fishery	0-1 (1)	1-1 (0)	1-1 (0)	2-1 (1)	4-1 (3)	0-1 (1)	6-1 (5)	4-1 (3)	0-1 (1)	7-1 (6)	21
South Coast Charter Angling	0-1 (1)	1-1 (0)	1-1 (0)	2-1 (1)	4-1 (3)	0-1 (1)	2-1 (1)	4-1 (3)	0-1 (1)	7-1 (6)	17
The Blackwater Valley Cycle Route	0-1 (1)	1-1 (0)	3-1 (2)	2-1 (1)	5-1 (4)	0-1 (1)	7-1 (6)	4-1 (3)	0-1 (1)	4-1 (3)	22
The East Cork –	0-1 (1)	1-1 (0)	3-1 (2)	2-1 (1)	5-1 (4)	0-1 (1)	6-1 (5)	4-1 (3)	0-1 (1)	4-1 (3)	21

Cork City to Youghal Cycle Route											
Youghal Deep Sea Angling	0-1 (1)	1-1 (0)	1-1 (0)	2-1 (1)	4-1 (3)	0-1 (1)	3-1 (2)	4-1 (3)	0-1 (1)	8-1 (7)	19
Youghal Golf Club	**5-10 (5)**	**5-10 (5)**	**9-10 (1)**	**7-10 (3)**	**8-10 (2)**	**7-10 (3)**	**9-10 (1)**	**9-10 (1)**	**9-10 (1)**	**10-10 (0)**	**23**

The study did not highlight any major shortcoming that would have required major changes prior to conduct a full comparative experiment.

Appendix V – Second Experiment (Open Domains)

A.1. Tasks

Please note that all tasks are voluntary. You may choose to skip questions without penalty and without providing a reason.

Task 1: Browsing around a selected activity – Content base Failte Ireland (adaptive)

Select one of the following four activities:

A) Golf
B) Cycling
C) Hiking
D) Fishing

1) Gather general information about the selected activity on the health and fitness website. Do this for at least 3 minutes. (Do not switch activity)

2) Related to the selected activity please answer following question:

Golf: According to the study, the death rate for golfers is how much lower than for non-golfers of the same sex, age and socioeconomic status?

Answer:

Cycling: How many stages of cycling fitness can be found in the content?

Answer:

Hiking: How much calories does Hiking burn in average in one hour?

Answer:

Fishing: What recommendation can be found in relation to bait?

Answer:

3) Browse to at least one of the three websites recommended as location for the selected activity.

4) Follow recommendations related to the activity selected and inform yourself about the local availability of activities.

5) Browse to the hotel site and note the link title of the three most relevant hotels in the link recommendation widget (that you think relates most to your current activity and location interest).

1)

2)

3)

6) Click on the link you think is most relevant and again note the title of the (for you) most relevant hotel recommendations.

1)

2)

3)

7) Browse to the Service by clicking on the wripl logo in the upper left corner. Here click on interests to see your cross-site browsing interests as a word cloud.

How relevant is the word cloud to your cross-site browsing interests?

Rate from 1-10 with 1 least relevant:

<u>Task 1: Browsing around a selected activity – Content base Failte Ireland (Non-Adaptive)</u>

Select one of the following four activities:

A) Golf
B) Cycling
C) Hiking
D) Fishing

1) Gather general information about the selected activity on the health and fitness website. Do this for at least 3 minutes. (Do not switch activity)

2) Related to the selected activity please answer following question:

Golf: According to the study, the death rate for golfers is how much lower than for non-golfers of the same sex, age and socioeconomic status?

Answer:

Cycling: How many stages of cycling fitness can be found in the content?

Answer:

Hiking: How much calories does Hiking burn in average in one hour?

Answer:

Fishing: What recommendation can be found in relation to bait?

Answer:

3) Browse to at least one of the three websites recommended as location for the selected activity.

4) Follow at least two links related to the activity selected and inform yourself about the local availability of activities.

5) Browse to the hotel site and use the inbuilt search function to search for the most relevant hotels in relation to your activity and location interest (e.g. Golf in Dingle).

6) Note the link title of the three most relevant hotels in the result list of the search.

1)

2)

3)

Task 2: Browsing around a selected activity – Mixed Content Failte Ireland and news content (Adaptive)

Select one of the following four activities although it must be a different one then used in Task 1:

A) Golf
B) Cycling
C) Hiking
D) Fishing

1) Gather general information about the selected activity on the health and fitness website. Do this for at least 3 minutes. (Do not switch activity)

2) Related to the selected activity please answer following question:

Golf: According to the study, the death rate for golfers is how much lower than for non-golfers of the same sex, age and socioeconomic status?

Answer:

Cycling: How many stages of cycling fitness can be found in the content?

Answer:

Hiking: How much calories does Hiking burn in average in one hour?

Answer:

Fishing: What recommendation can be found in relation to bait?

Answer:

3) Browse to at least one of the three websites recommended as location for the selected activity.

4) Follow recommendations related to the activity selected and inform yourself about the local availability of activities.

5) Click on the link you think is most relevant and again note the title of the link recommendations that you think are most relevant for you.

1)

2)

3)

6) Click on the link you think is most relevant and again note the title of the (for you) most relevant news recommendations.

1)

2)

3)

7) Browse to the Service by clicking on the wripl logo in the upper left corner. Here click on interests to see your cross-site browsing interests as a word cloud.

How relevant is the word cloud to your cross-site browsing interests?

Rate from 1-10 with 1 least relevant:

Task 2: Browsing around a selected activity – Mixed Content base Failte Ireland and News Content (Non-Adaptive)

Select one of the following four activities although it must be a different one then used in Task 1:

A) Golf
B) Cycling
C) Hiking
D) Fishing

1) Gather general information about the selected activity on the health and fitness website. Do this for at least 3 minutes. (Do not switch activity)

2) Related to the selected activity please answer following question:

Golf: According to the study, the death rate for golfers is how much lower than for non-golfers of the same sex, age and socioeconomic status?

Answer:

Cycling: How many stages of cycling fitness can be found in the content?

Answer:

Hiking: How much calories does Hiking burn in average in one hour?

Answer:

Fishing: What recommendation can be found in relation to bait?

Answer:

3) Browse to at least one of the three websites recommended as location for the selected activity.

4) Follow links related to the activity selected and inform yourself about the local availability of activities.

5) Browse to the news site and use the inbuilt search function to search for the most relevant articles in relation to your activity and location interest (e.g. Golf in Dingle).

6) Note the link title of the three most relevant articles in the result list of the search.

1)

2)

3)

A.2. Task performance metrics (through logging)

User Action Logging

1. Task completion time
2. Number of queries per task

3. Number of clicks to answer question
4. Number of pages viewed
5. Number of mouse movements

User Information Foraging

6. Amount of collected information
7. Diversity of collected information

User Sense Making

8. Task performance (answer accuracy & completeness)
9. Ratio of successes to failures

A.3. Pre-test Questionnaire

Please note that all questions are voluntary. You may choose to skip questions without penalty and without providing a reason.

	Strongly disagree				Strongly agree
1. I browse the web on a regular base.	1	2	3	4	5
2. I follow link recommendations, e.g. on the bottom of the Amazon website on a regular base.	1	2	3	4	5
	1	2	3	4	5
3. I follow advertised links, (e.g. on top of Googles search result) on a regular base.	1	2	3	4	5
4. I have used adaptive/personalised web systems in the past (e.g. Googles personalised search).	1	2	3	4	5
	1	2	3	4	5
5. I have concerns about the usage of my personal browsing information on a regular base.	1	2	3	4	5
	1	2	3	4	5
	1	2	3	4	5
	1	2	3	4	5

A.4. Post Questionnaires

Questionnaire Scale 0-100

0 strongly disagree
25 Disagree
50 neutral
75 agree
100 strongly agree

For yes/no questions

0 - yes and 1 - no

1. How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? In other words, was the task easy (low mental demand) or complex (high mental demand)?
2. How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely (low temporal demand) or rapid and frantic (high temporal demand)?
3. How secure and relaxed (low stress) versus insecure and irritated (high stress) did you feel during the task?
4. How much conscious mental effort or concentration was required?
5. How successful do you think you were in accomplishing the goal of the task? How satisfied were you with your performance in accomplishing the goal?
6. How often did interruptions on the task occur? In other words, were distractions (mobile, questions, noise, etc.) not important (low context bias) or did they influence your task (high context bias)?
7. How much attention was required for activities like remembering, problem-solving, decision-making and perceiving (eg. detecting, recognizing and identifying objects)?
8. How much attention was required for spatial processing (to pay attention around you)?
9. How much attention was required for verbal material (eg. reading or processing linguistic material)?

10. How much attention was required for executing the task based on the information visually received (through eyes)?
11. How much attention was required to manually respond to the task? In other words, did you need to pay low or high attention to using the keyboard/mouse?
12. How much experience do you have in performing the task or similar tasks on the Web sites used in the experiment?
13. To what extent did your internet/web browsing skills facilitate the execution of the task?
14. Were you motivated to complete the task?
15. Did you perform just this task (low parallelism) or were you doing other parallel tasks (high parallelism) (e.g. multiple tabs/windows/programs)?
16. Were you sleepy, tired (low arousal) or fully awake and activated (high arousal)?
17. It was difficult to find information related to the task when browsing across the separate Web sites.
18. Do you know if you have ever used Web sites that recommend content relevant to your task or interest?
19. I think a Web site that recommends relevant content based on my browsing behaviour is useful.
20. I would have privacy concerns if a Web site would recommend content based on my browsing behaviour.
21. The system guided me towards relevant content.
22. I think the task reflected a real world browsing task.

A.5. Concluding questions and comments to overall personalisation approach

Please note that all questions are voluntary. You may choose to skip questions without penalty and without providing a reason.

	Strongly Disagree				Strongly Agree
I found the adaptive hints were relevant related to my cross-site browsing task.	1	2	3	4	5
I found the presentation of the adaptive hints was too intrusive.	1	2	3	4	5
I found the presentation of the adaptive hints distracting.	1	2	3	4	5
The adaptive hints did not help me solve the tasks.	1	2	3	4	5
The adaptive hints motivated me to explore information related to the task.	1	2	3	4	5

Concluding Comments:

Findings User Performance

Click related

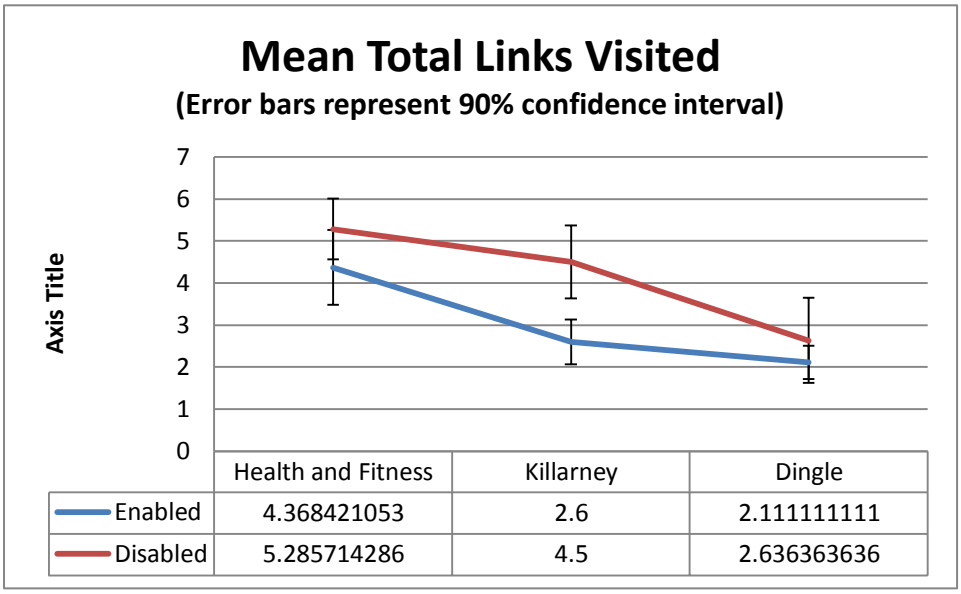


Figure 63 Mean Total Links Visited All Websites

System Setting	90% Conf. Int. Health and Fitness	90% Conf. Int. Killarney	90% Conf. Int. Dingle
Enabled	0.891	0.538	0.395
Disabled	0.716	0.968	1.016

Table 8 90% Confidence Interval Total Links Visited

System Setting	95% Conf. Int. Health and Fitness	95% Conf. Int. Killarney	95% Conf. Int. Dingle
Enabled	1.08	0.651	0.479
Disabled	0.865	1.049	1.228

Table 9 95% Confidence Interval Total Links Visited

In relation to the overall clicks the overall mean clicks is overall less in the personalised system setting with 10.42 clicks vs. 14.3 clicks.

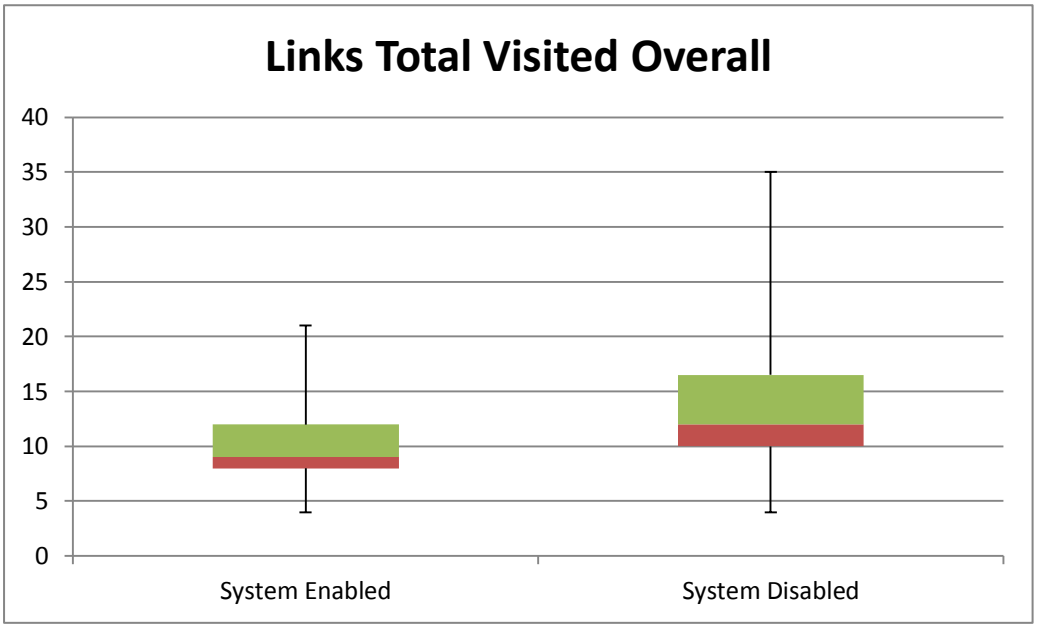


Figure 64 Box chart Total Links Visited

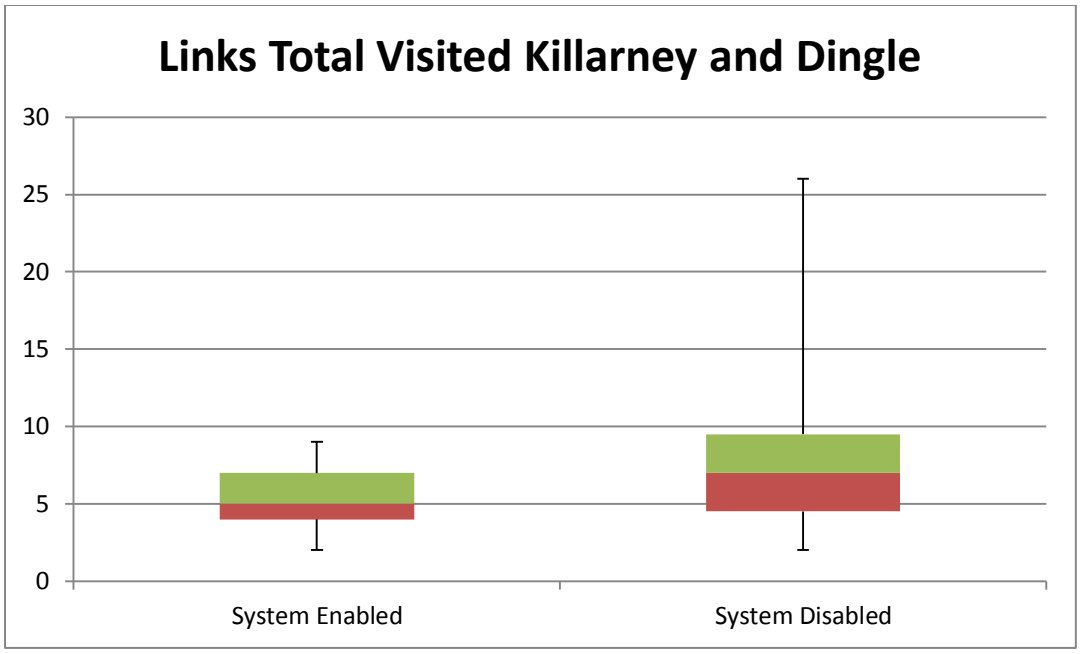


Figure 65 Box chart Total Links Vi sited Killarney and Dingle

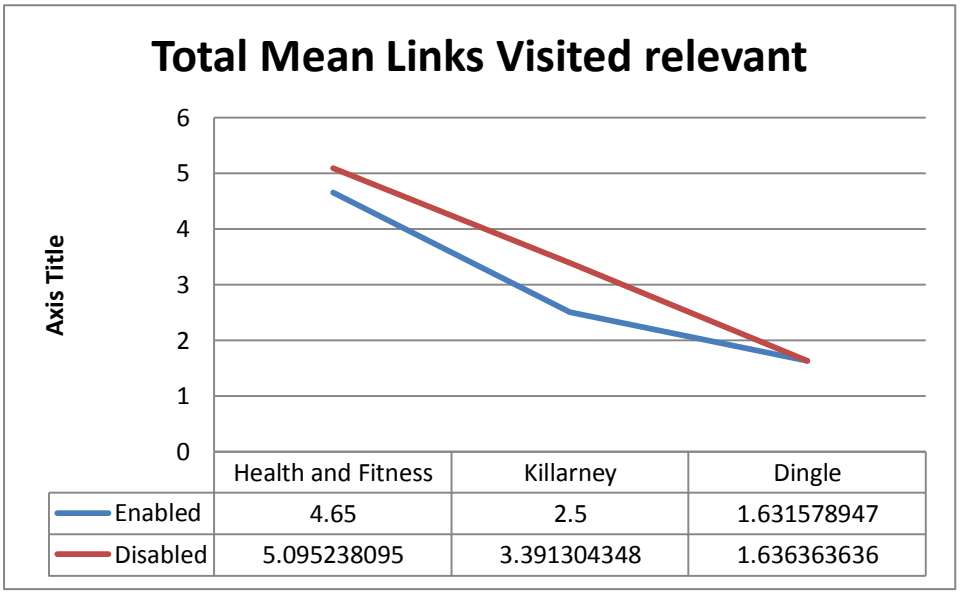


Figure 66 Line Chart Average Relevant Links Visited

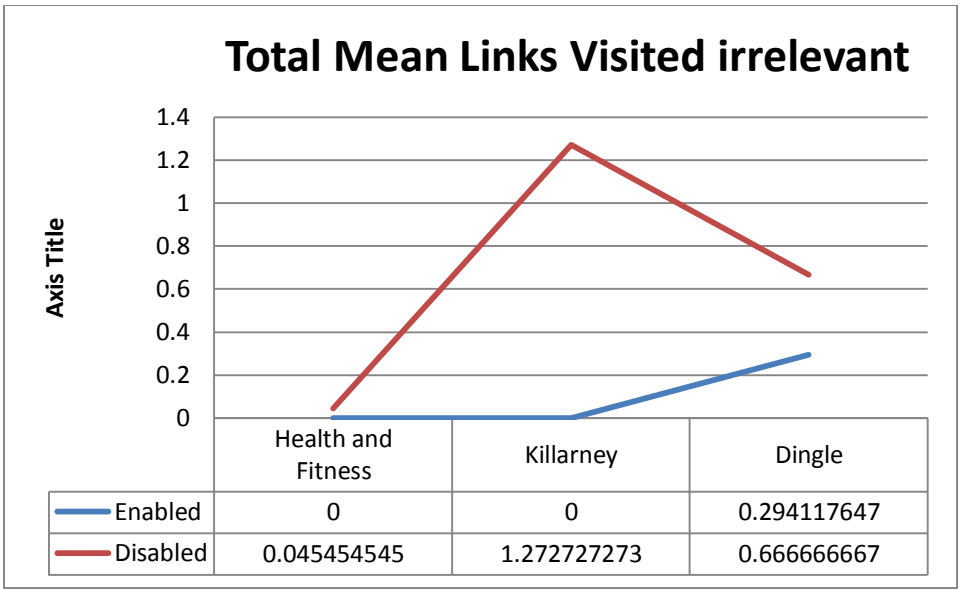


Figure 67 Line Chart Average Irrelevant Links Visited

Task Completion Time

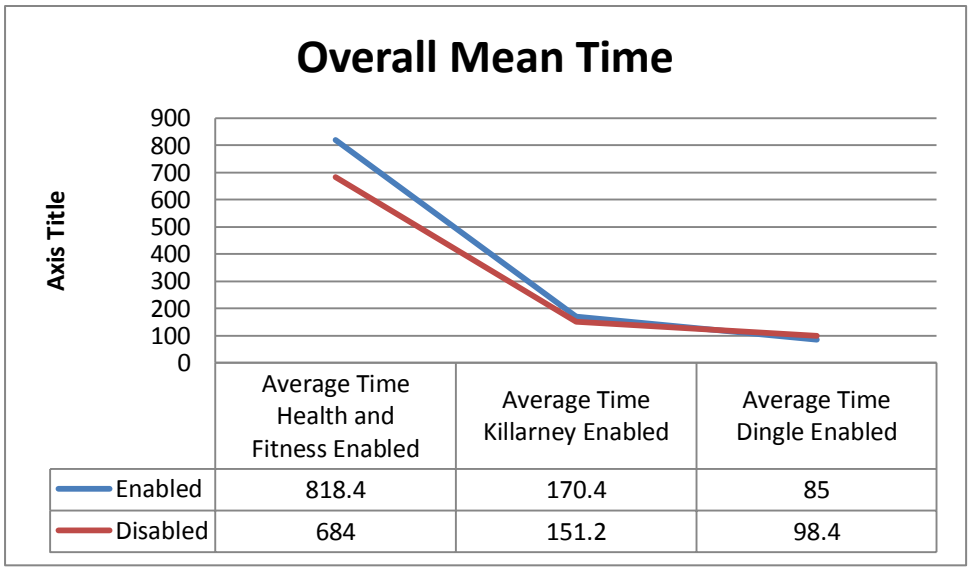


Figure 68 Overall Mean Completion Time All Websites

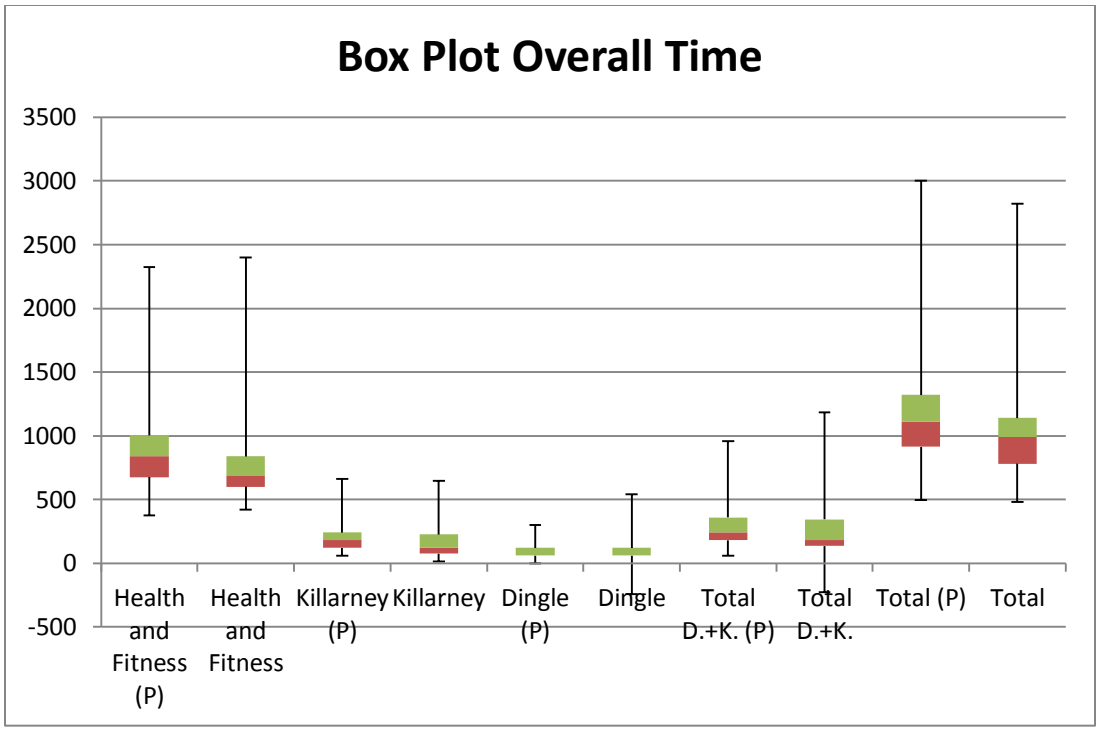


Figure 69 Box Plot Aggregation of Overall Time All Websites

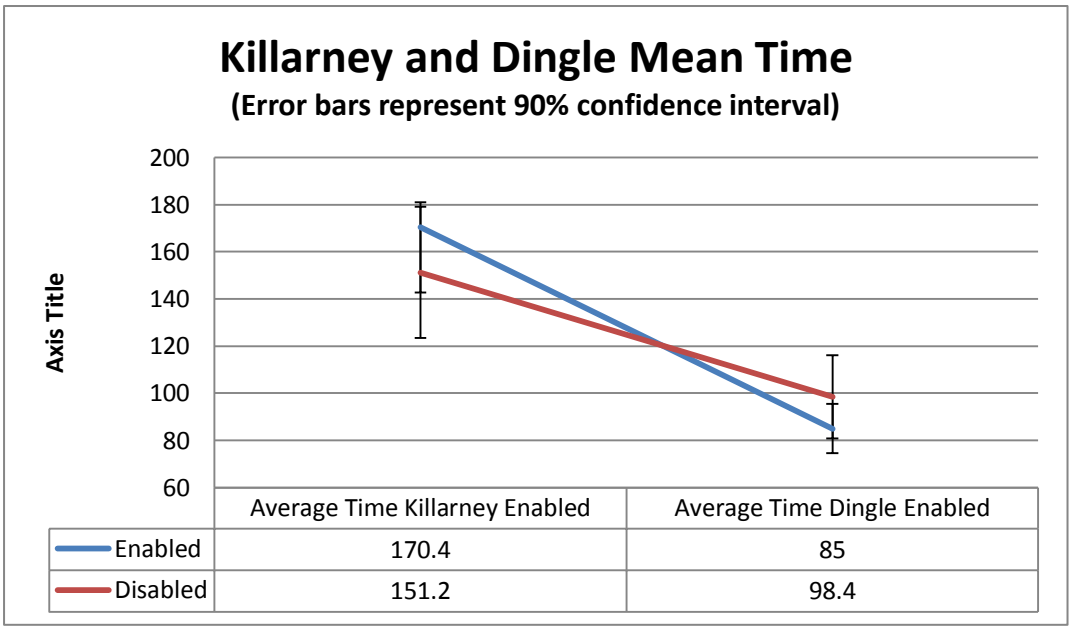


Figure 70 Line Chart Killarney and Dingle Average Overall Time

System Setting	90% Conf. Int. Killarney	90% Conf. Int. Dingle
Enabled	27.60	10.57
Disabled	27.87	17.66

Table 10 90% Confidence Interval Average Time Dingle and Killarney

System Setting	90% Conf. Int. Killarney	90% Conf. Int. Dingle
Enabled	33.29	12.76
Disabled	33.63	21.31

Table 11 95% Confidence Interval Average Time Dingle and Killarney

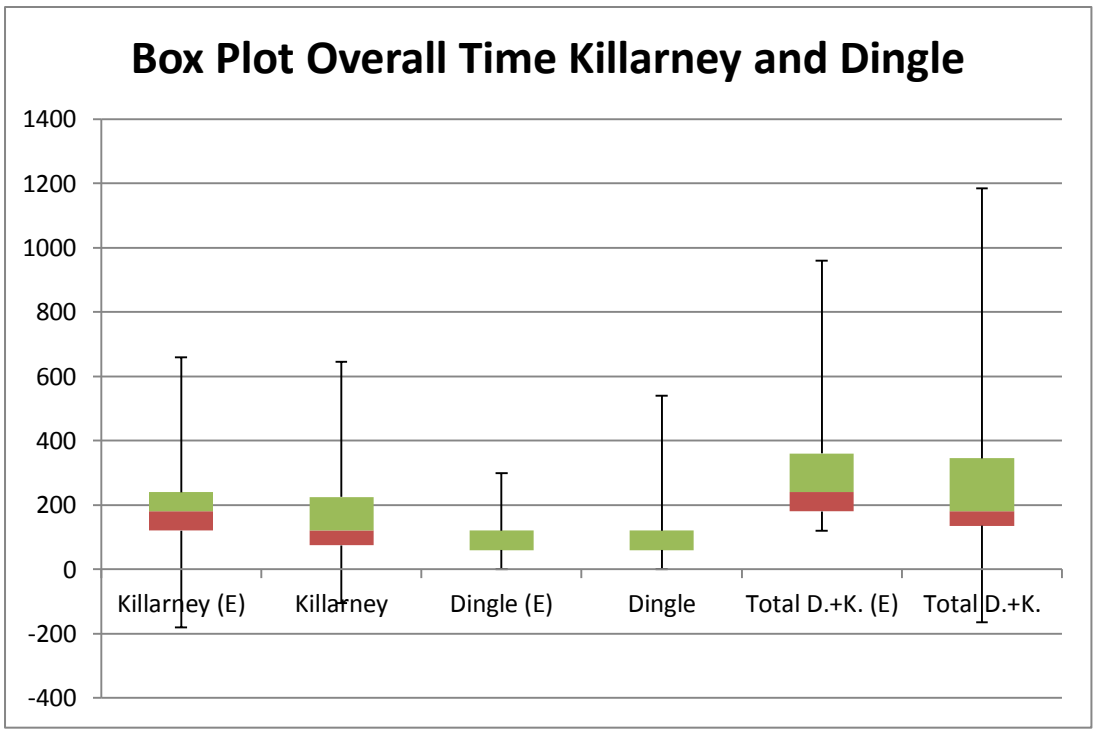


Figure 71 Boxplot Overview Overall Time Killarney and Dingle

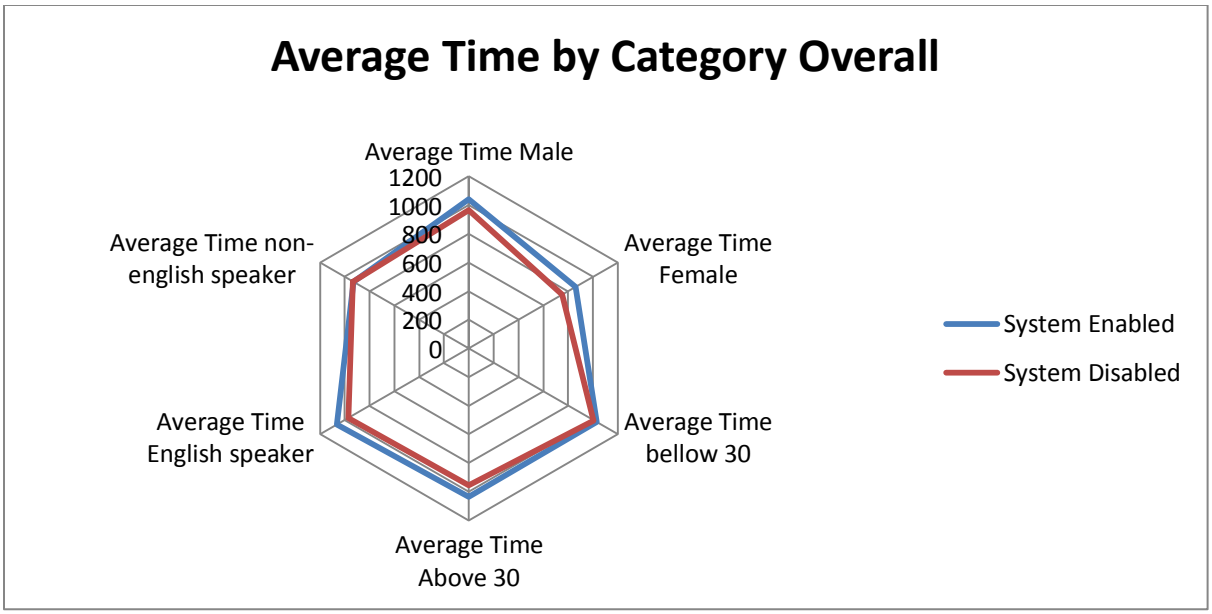


Figure 72 Average Time per Category Overall

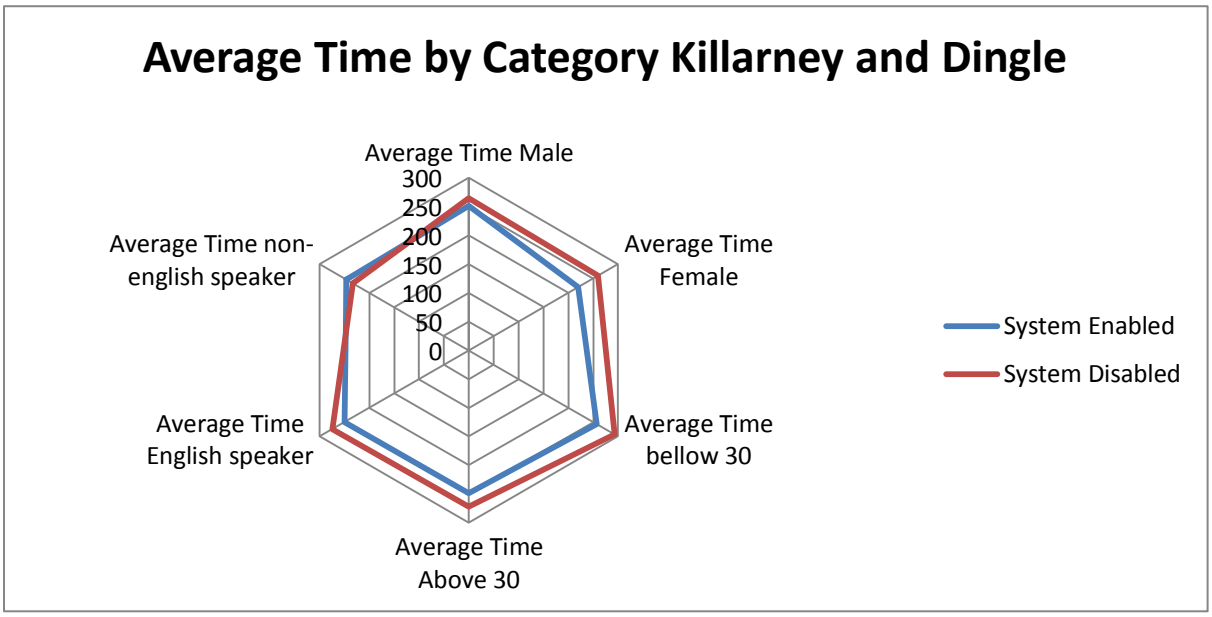


Figure 73 Average Time per Category Killarney Dingle

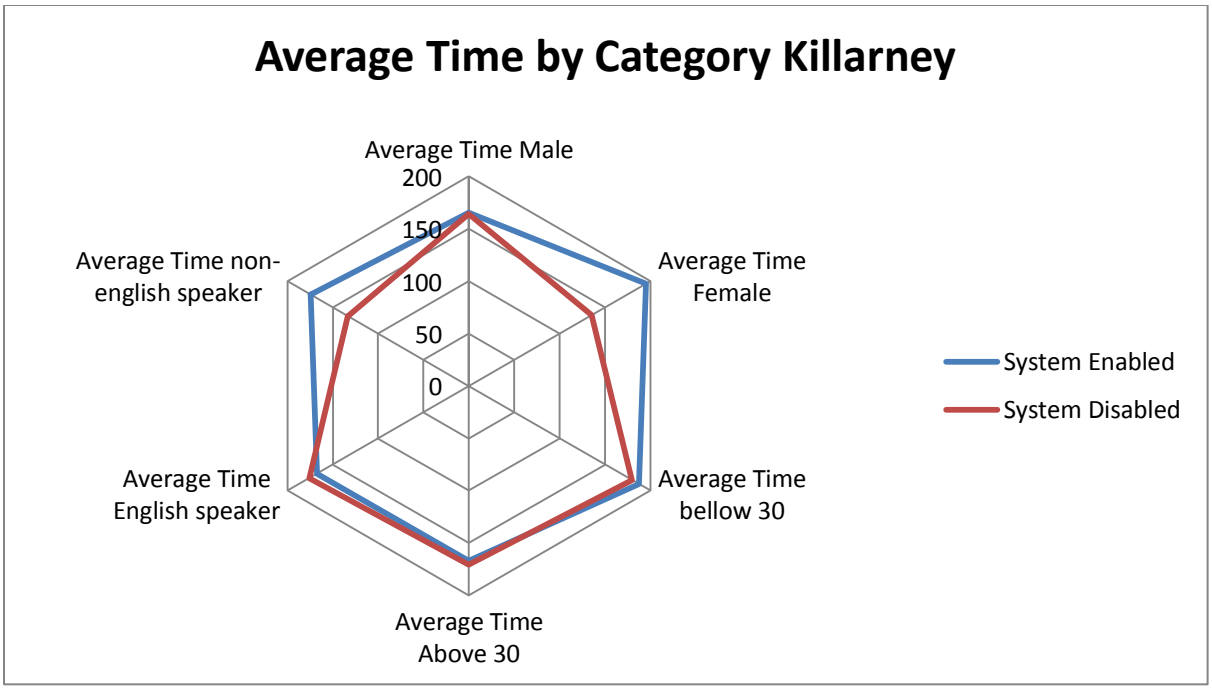


Figure 74 Average Time per Category Killarney

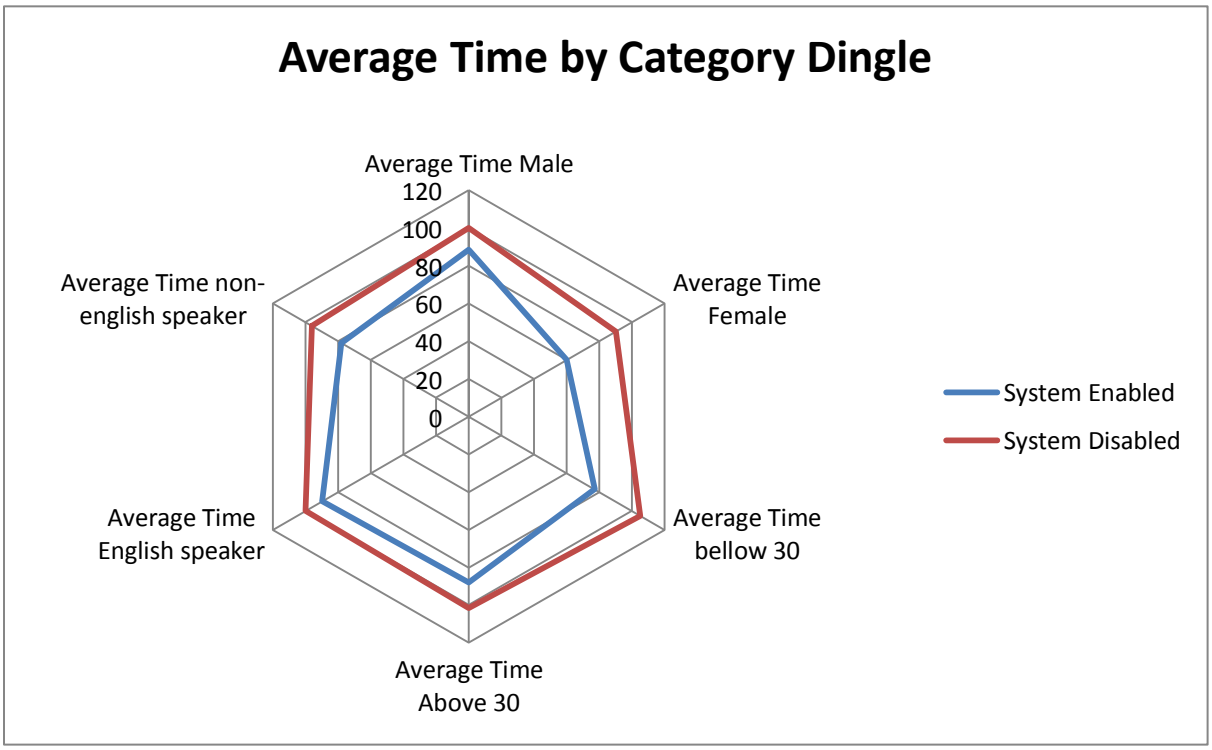


Figure 75 Average Time per Category Dingle

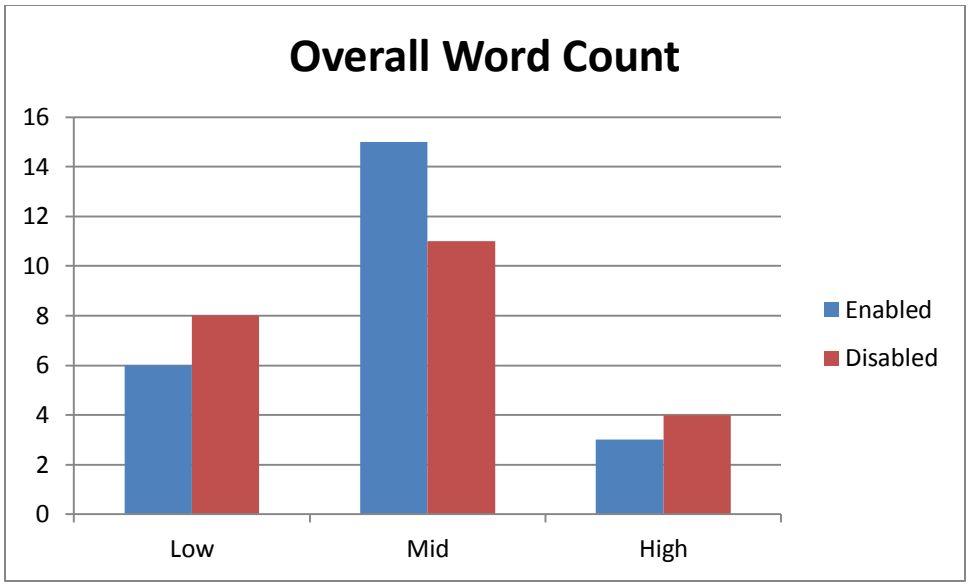


Figure 76 Overall Word Count

Findings Self-Reported Feedback

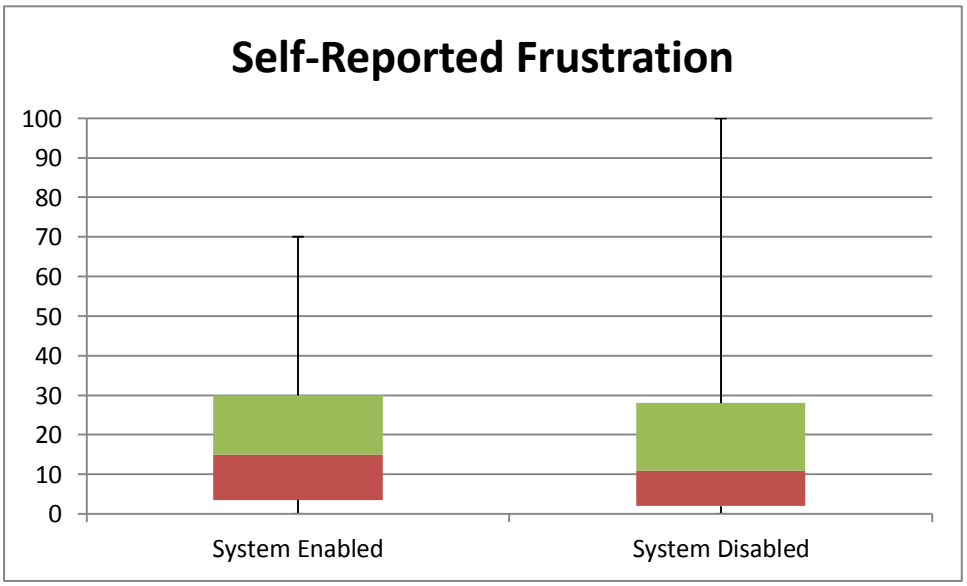


Figure 77 Self-Reported Frustration Second Experiment

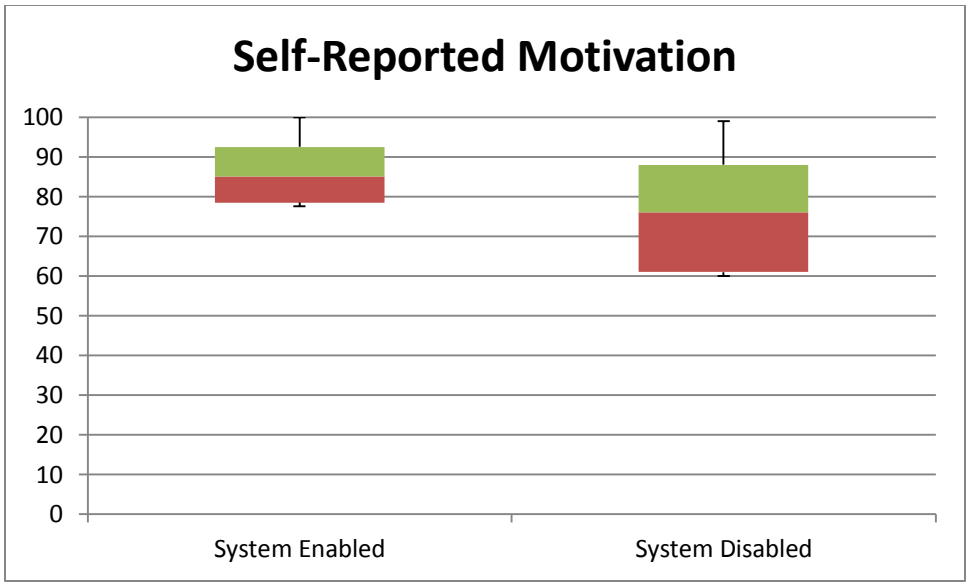


Figure 78 Self-Reported Motivation Second Experiment

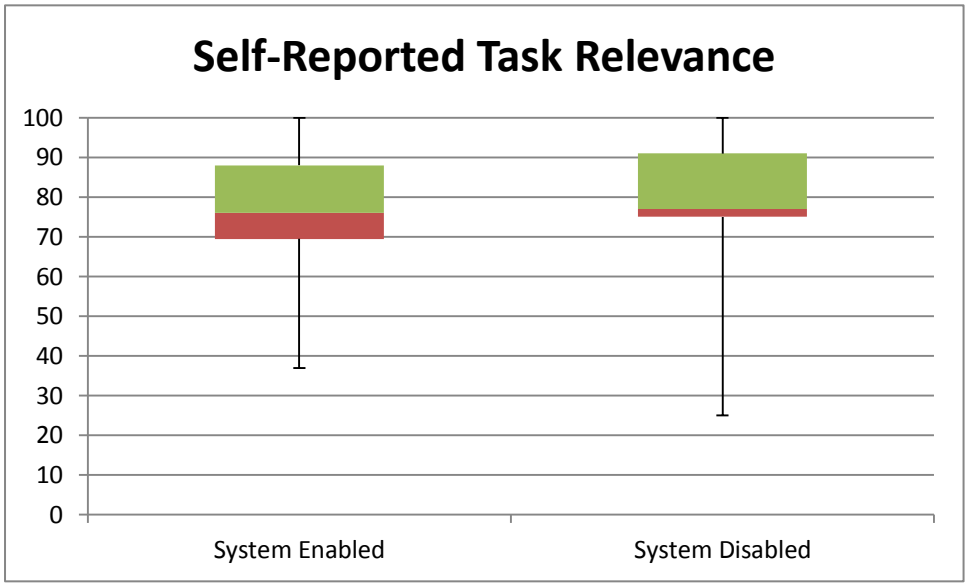


Figure 79 Self-Reported Task Relevance Second Experiment

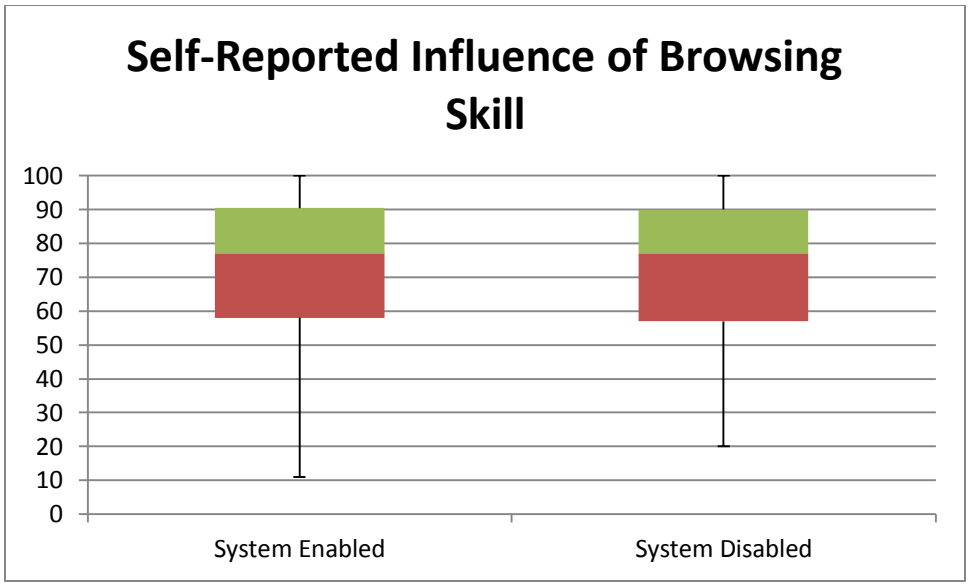


Figure 80 Influence of Browsing Skill Second Experiment

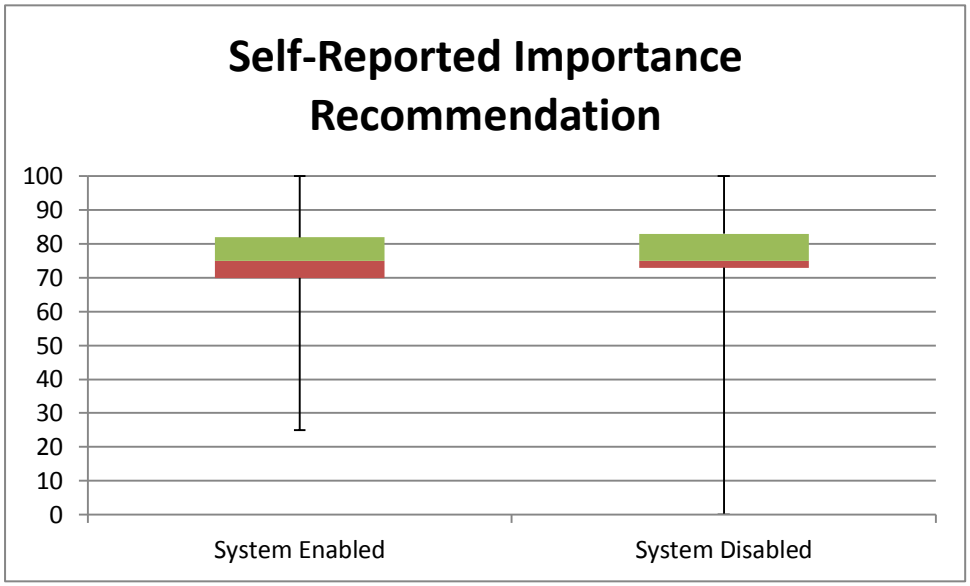


Figure 81 Self-Reported Importance of the Recommendations Second Experiment

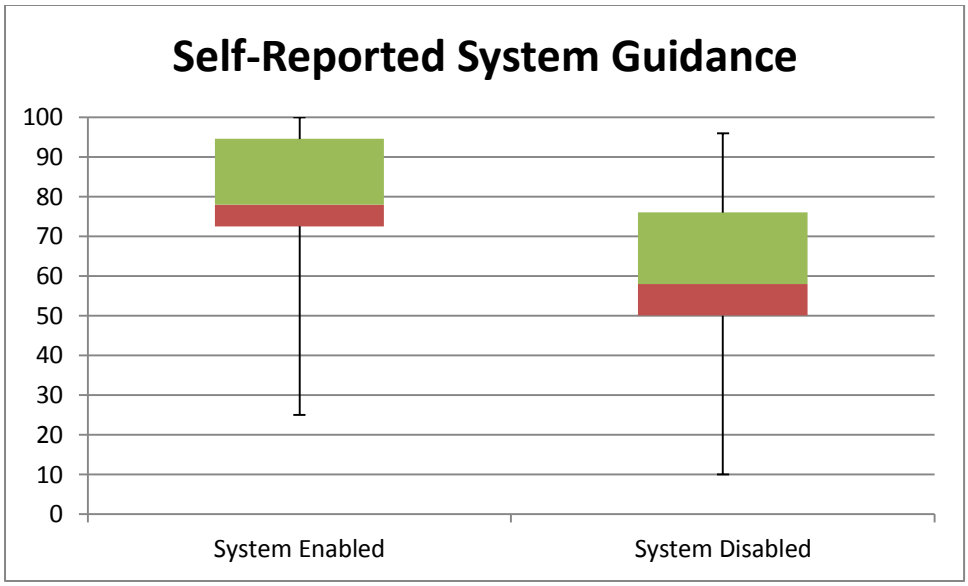
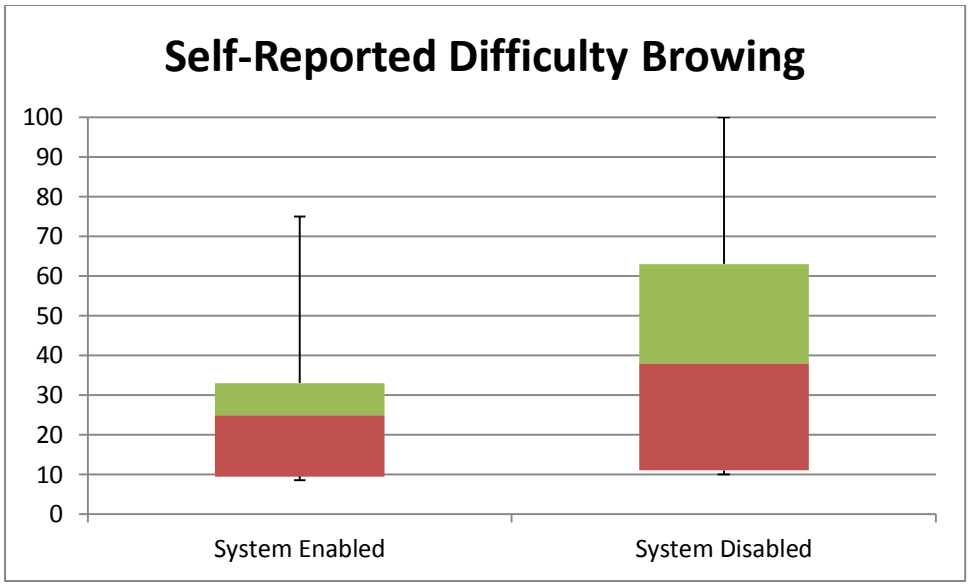
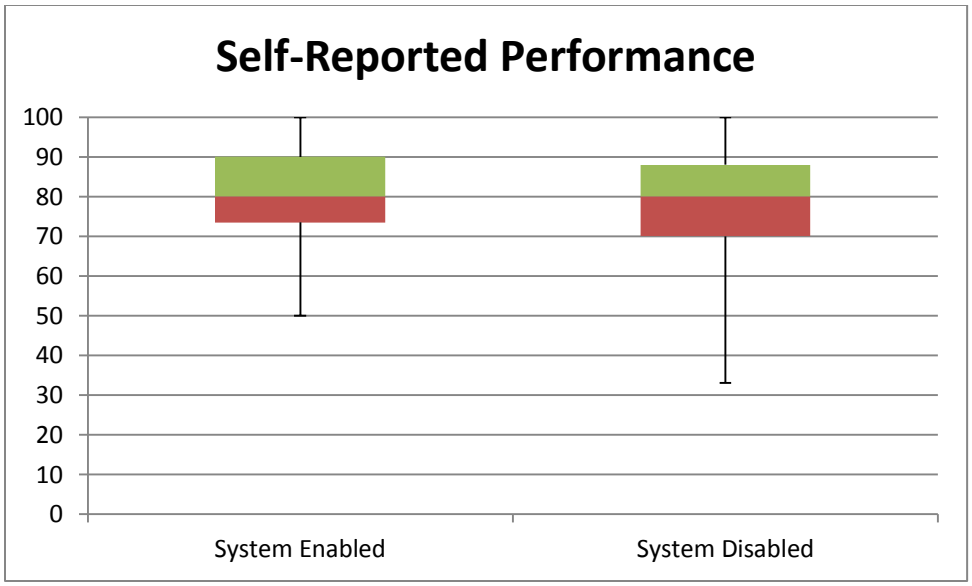


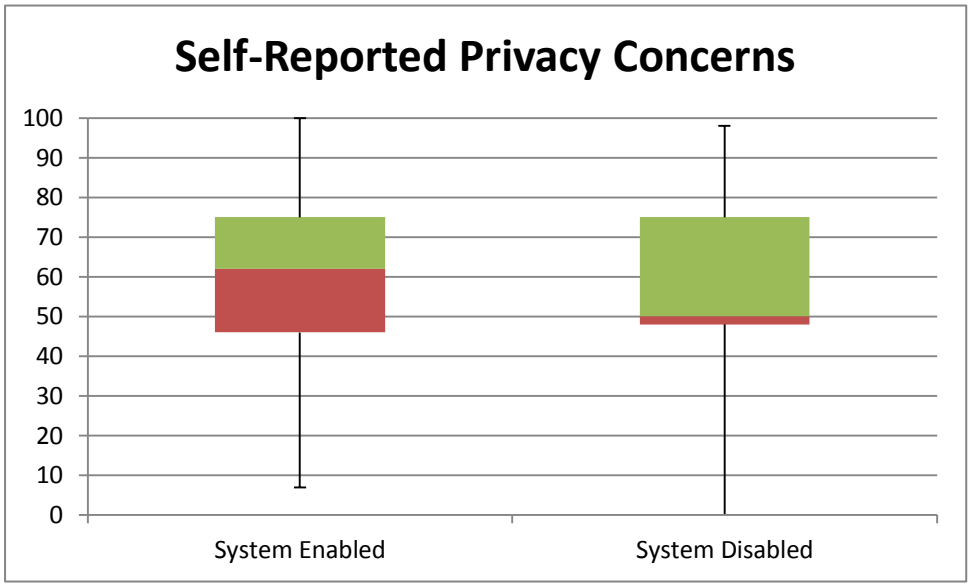
Figure 82 Self-Reported Usefulness of System Guidance Second Experiment

t-Test: Two-Sample Assuming Equal Variances		
	system_guidance_q25_enabled	system_guidance_q25_enabled
Mean	79.55555556	59.4137931
Variance	333.1794872	521.7512315
Observations	27	29
Pooled Variance	430.9574287	
Hypothesized Mean Difference	0	
df	54	
t Stat	3.62799894	
P(T<=t) one-tail	0.00031717	
t Critical one-tail	1.673564906	
P(T<=t) two-tail	0.000634341	
t Critical two-tail	2.004879288	



t-Test: Two-Sample Assuming Equal Variances		
	difficulty_browsing_q20_enabled	difficulty_browsing_q20_disabled
Mean	20.28	34.25
Variance	181.96	705.9722222
Observations	25	28
Pooled Variance	459.3782353	
Hypothesized Mean Difference	0	
df	51	
t Stat	-2.368766671	
P(T<=t) one-tail	0.010836581	
t Critical one-tail	1.67528495	
P(T<=t) two-tail	0.021673161	
t Critical two-tail	2.00758377	


Self-Reported Performance
100
90
80
70
60
50
40
30
20
10
0
System Enabled
System Disabled


Self-Reported Privacy Concerns
100
90
80
70
60
50
40
30
20
10
0
System Enabled
System Disabled

Integration Overview of CSP in Open Domains Part 1 (Figure is split into two parts)

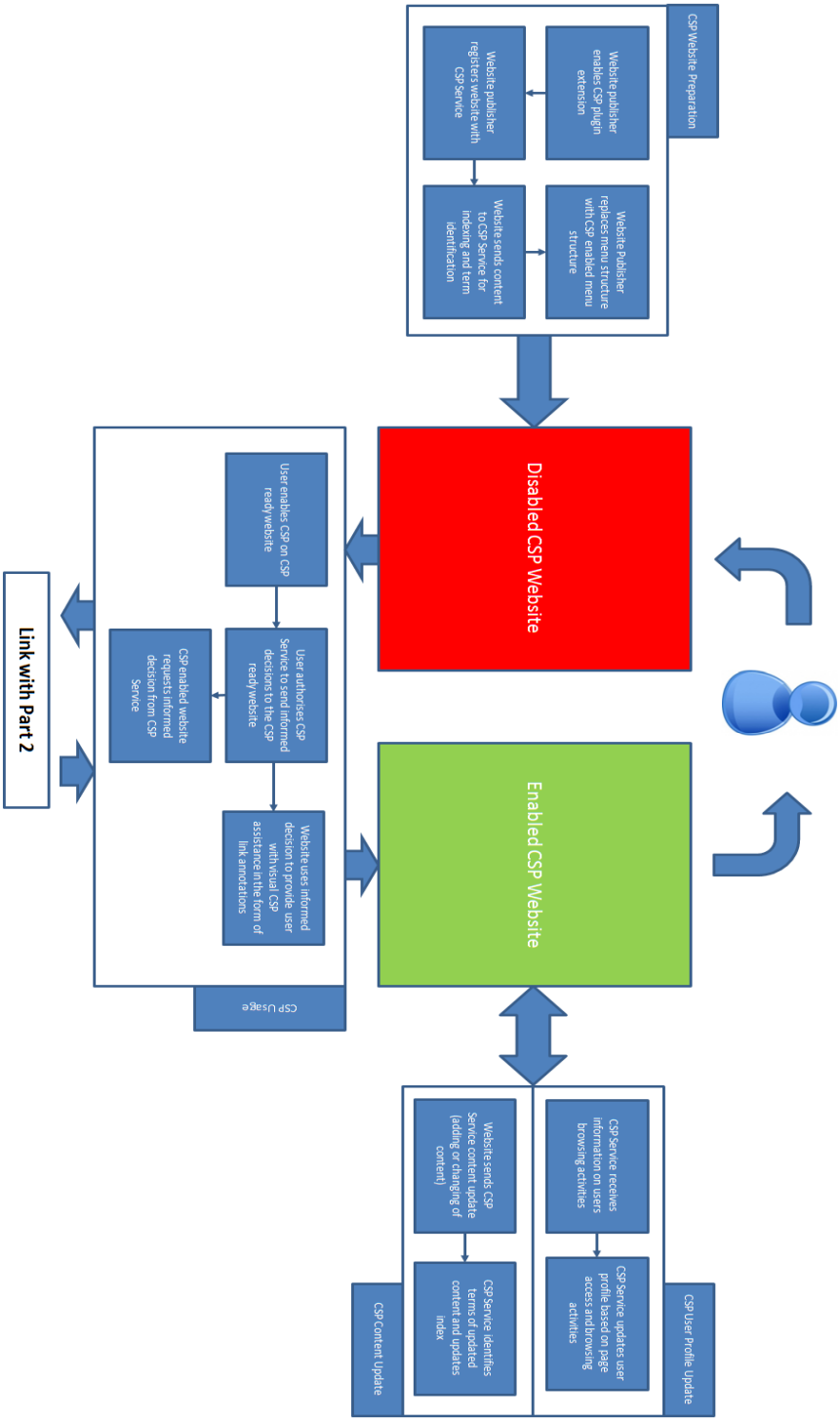


Figure 83 Integration Overview of CSP in Open Domains Part 1

Integration Overview of CSP in Open Domains Part 2 (Figure is split into two parts)

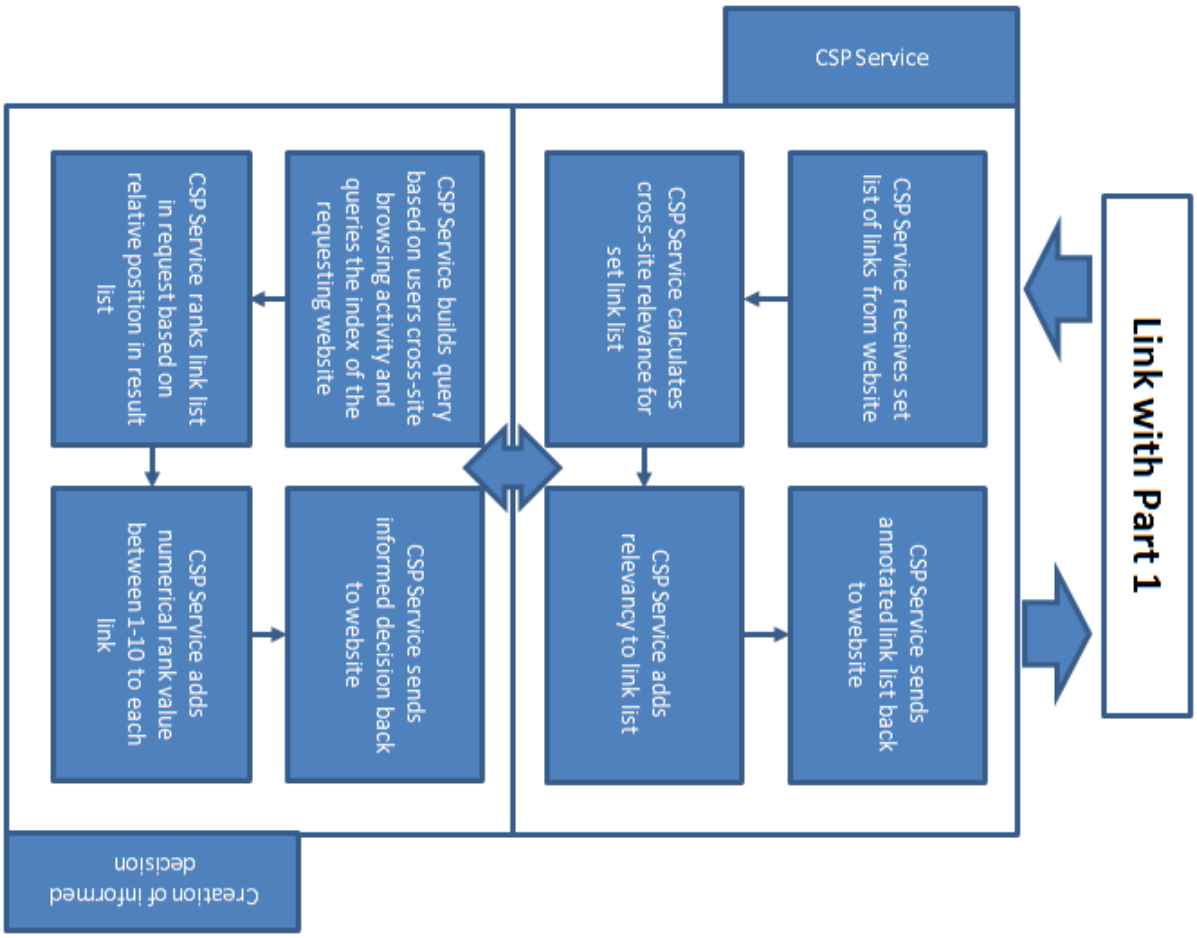


Figure 84 Integration Overview of CSP in Open Domains Part 2