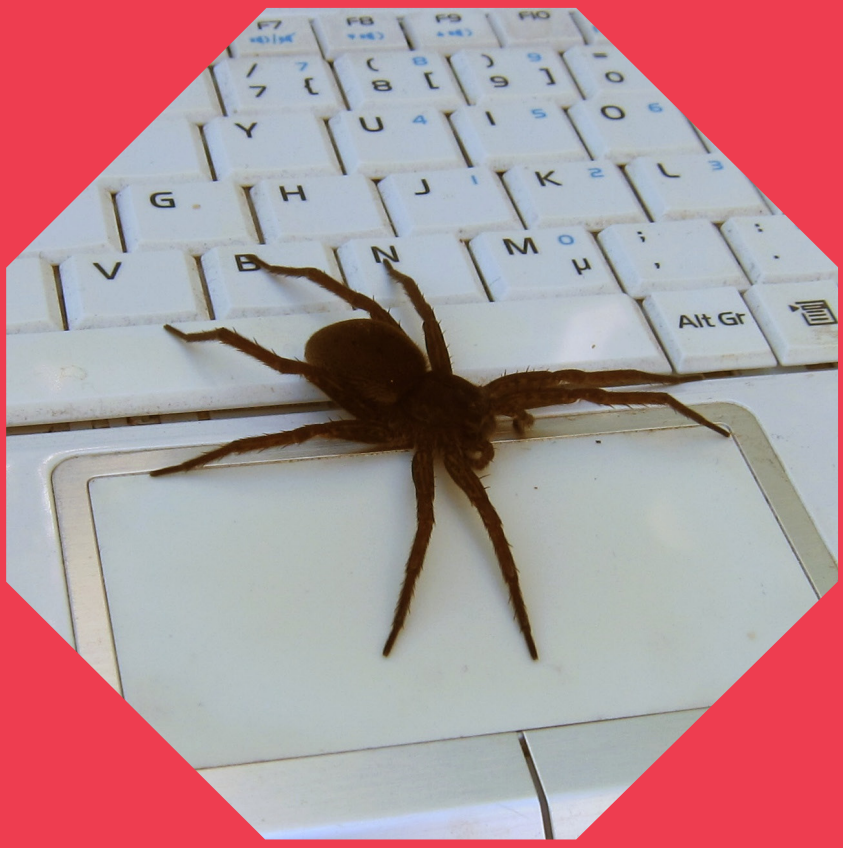
Department of Media Technology

Methods for Building Semantic Portals

Osma Suominen



Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 113/2013

Methods for Building Semantic Portals

Osma Suominen

A doctoral dissertation completed for the degree of Doctor of Technology to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall TU2 of the school on 9 September 2013 at 12.

Aalto University
School of Science
Department of Media Technology
Semantic Computing Research Group

Supervising professor
Eero Hyvönen

Preliminary examiners
Prof. Vagan Terziyan, University of Jyväskylä, Finland
Dr. Antoine Isaac, Vrije Universiteit Amsterdam, the Netherlands

Opponent
Dr. Thomas Baker, Dublin Core Metadata Initiative, United States

Aalto University publication series
DOCTORAL DISSERTATIONS 113/2013

ISBN 978-952-60-5253-3 (printed)
ISBN 978-952-60-5254-0 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-5254-0

Unigrafia Oy
Helsinki 2013

Finland

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Osma Suominen

Name of the doctoral dissertation
Methods for Building Semantic Portals

Publisher School of Science

Unit Department of Media Technology

Series Aalto University publication series DOCTORAL DISSERTATIONS 113/2013

Field of research Computer Science, Semantic Web

Manuscript submitted 9 April 2013 **Date of the defence** 9 September 2013

Permission to publish granted (date) 31 May 2013 **Language** English

☐ **Monograph** ☒ **Article dissertation (summary + original articles)**

Abstract

Semantic portals are information systems which collect information from several sources and combine them using semantic web technologies into a user interface that solves information needs of users. Creating such portals requires methods and tools from multiple disciplines, including knowledge representation, information retrieval, information extraction, and user interface design.

This thesis explores methods for building and improving semantic portals and other semantic web applications with contributions in three areas. The studies included in the thesis draw from the design science methodology in information systems research.

First, a method for creating of faceted search user interfaces for semantic portals utilizing controlled vocabularies with a complex hierarchical structure is presented. The results show that the method allows the creation of user-centric search facets that hide the complex hierarchies from the user, resulting in a user-friendly faceted search interface.

Second, the creation of structured metadata from text documents is enhanced by adapting a state of the art automatic subject indexing system to Finnish language texts. The results show that using a suitable combination of existing tools, automatic subject indexing quality comparable to that of human indexers can be attained in a highly inflected language such as Finnish.

Finally, the quality of controlled vocabularies such as thesauri and lightweight ontologies is examined by developing a set of quality criteria for vocabularies expressed using the SKOS standard, and methods for correcting structural problems in SKOS vocabularies are presented. The results show that most published SKOS vocabularies suffer from quality issues and violate the SKOS integrity conditions. However, the great majority of such problems were corrected by the methods presented in this dissertation.

The methods have been implemented in several real world applications, including the HealthFinland health information portal, the ARPA information extraction toolkit, and the ONKI ontology library system.

Keywords semantic web, faceted search, automatic subject indexing, vocabulary quality

ISBN (printed) 978-952-60-5253-3 **ISBN (pdf)** 978-952-60-5254-0

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2013

Pages 161 **urn** http://urn.fi/URN:ISBN:978-952-60-5254-0

Aalto-yliopisto

Tiivistelmä

Aalto-yliopisto, PL 11000, 00076 Aalto www.aalto.fi

Tekijä
Osma Suominen

Väitöskirjan nimi
Methods for Building Semantic Portals

Julkaisija Perustieteiden korkeakoulu

Yksikkö Mediatekniikan laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 113/2013

Tutkimusala tietojenkäsittelytiede, semanttinen web

Käsikirjoituksen pvm 09.04.2013 **Väitöspäivä** 09.09.2013

Julkaisuluvan myöntämispäivä 31.05.2013 **Kieli** Englanti

☐ **Monografia** ☒ **Yhdistelmäväitöskirja (yhteenveto-osa + erillisartikkelit)**

Tiivistelmä

Semanttiset portaalit ovat tietojärjestelmiä, jotka keräävät tietoa useista lähteistä ja yhdistävät ne semanttisen webin teknologioiden avulla käyttäjien tiedontarpeita tukevaksi käyttöliittymäksi. Tällaisten portaalien rakentaminen vaatii menetelmiä ja työkaluja useilta tieteenaloilta, mukaan lukien tietämyksen esittäminen, tiedonhaku, tiedon eristäminen ja käyttöliittymäsuunnittelu.

Tässä väitöskirjassa tarkastellaan menetelmiä semanttisten portaalien ja muiden semanttisen webin sovellusten rakentamiseksi. Väitöskirjan tulokset jakaantuvat kolmeen osa-alueeseen. Tutkimuksessa käytetyt menetelmät perustuvat informaatiojärjestelmien tutkimuksessa käytettyihin suunnittelutieteen menetelmiin.

Ensiksi väitöskirjassa esitetään menetelmä semanttisten portaalien fasettipohjaisten käyttöliittymien luomiseksi monimutkaisten kontrolloitujen sanastojen pohjalta. Tulokset osoittavat, että menetelmä mahdollistaa sellaisten käyttäjäkeskeisten hakunäkymien luomisen, jotka piilottavat monimutkaiset hierarkiat käyttäjältä ja auttavat siten luomaan käyttäjäystävällisen fasettipohjaisen hakukäyttöliittymän.

Toiseksi rakenteisen metatiedon tuottamista tekstidokumenteista parannetaan sovittamalla nykyaikainen automaattisen sisällönkuvailun järjestelmä suomenkieliselle tekstiaineistolle. Tulokset osoittavat, että käyttämällä sopivaa yhdistelmää olemassaolevista työkaluista saavutetaan ihmistyönä tehtyyn sisällönkuvailuun verrattavissa oleva automaattisen sisällönkuvailun laatu myös agglutinatiivisella kielellä kuten suomen kielellä esitetyille aineistoille.

Kolmanneksi tarkastellaan kontrolloitujen sanastojen kuten asiasanastojen ja kevytontologioiden laatua kehittämällä laatukriteeristö SKOS-standardin avulla esitetyille sanastoille sekä esittämällä menetelmiä SKOS-sanastojen rakenteisten ongelmien korjaamiseksi. Tulokset osoittavat, että useimmat julkaistut SKOS-sanastot kärsivät laatuongelmista eivätkä noudata SKOS-standardin eheyssääntöjä. Suuri osa näistä ongelmista pystyttiin korjaamaan tässä väitöskirjassa esitetyin menetelmin.

Menetelmät on toteutettu useissa käytössä olevissa järjestelmissä, kuten TerveSuomi-terveystietoportaalissa, ARPA-tiedoneristämistyökalussa ja ONKI-ontologiakirjastossa.

Avainsanat semanttinen web, fasettihaku, automaattinen sisällönkuvailu, sanastojen laatu

ISBN (painettu) 978-952-60-5253-3 **ISBN (pdf)** 978-952-60-5254-0

ISSN-L 1799-4934 **ISSN (painettu)** 1799-4934 **ISSN (pdf)** 1799-4942

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2013

Sivumäärä 161 **urn** http://urn.fi/URN:ISBN:978-952-60-5254-0

Preface

The research presented in this dissertation was carried out at the Semantic Computing Research Group (SeCo) at the Department of Media Technology, Aalto University (formerly the Helsinki University of Technology) and the Department of Computer Science, University of Helsinki, Finland.

I warmly thank my supervisor, Professor Eero Hyvönen, who has provided all the support, guidance, and inspiration that made this research not only possible but also very enjoyable. I also appreciate the comments and suggestions for improvements made by the pre-examiners of this dissertation, Professor Vagan Terziyan and Antoine Isaac.

I also thank my co-workers and other researchers at SeCo and elsewhere, who have contributed to this dissertation either directly or indirectly. Kim Viljanen has been a source of joy, inspiration and enthusiasm as well as plenty of lateral thinking. Eetu Mäkelä challenged me to get involved in Semantic Web research, and I am thankful for our many valuable and fruitful discussions during the years. Tomi Kauppinen, Nina Laurenne, and Tuukka Ruotsalo have all set examples in sticking to scientific ideals in a challenging research environment with sometimes conflicting goals. I also thank Miika Alonen, Matias Frosterus, Markus Holi, Alex Johansson, Jussi Kurki, Joonas Laitio, Aleksi Lindblad, Petri Lindgren, Sini Pessala, Katri Seppälä, Reetta Sinkkilä, Jouni Tuominen, Juha Törnroos, Henri Ylikotila, and everyone else at SeCo I forgot to mention, for all the fruitful collaboration we have had and, in many cases, continue to have. Finally, I thank Christian Mader for the opportunity to combine our research efforts and for all the flexibility and persistence that it took to finish our joint work.

The research has been funded mainly by Tekes, the Finnish Funding Agency for Technology and Innovation. I am grateful for the support and research opportunities provided by the participating organizations in the

Tekes projects, particularly the Finnish Institute for Health and Welfare (THL), the Finnish Society for Social and Health (SOSTE), the Ministry of Employment and the Economy (TEM), and the National Library of Finland. My warm thanks go to Eija Hukka, who worked with inspiration and dedication on the HEALTHFINLAND project. I also thank Johanna Eerola for our long collaboration on health ontologies, and Hanna Heikkonen and Soila Veltheim for the opportunity to work with Sosiaaliportti content.

Finally, I thank my parents and my brother for setting me off on the path that lead to this dissertation, and my family Kaisa, Kerttu and Ilmari for all their unending love, support and understanding and for giving meaning to my life.

Evitskog, July 2, 2013,

Osma Suominen

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Osma Suominen, Kim Viljanen and Eero Hyvönen. User-centric Faceted Search for Semantic Portals. In *Proceedings of the 4th European Semantic Web Conference (ESWC2007)*, The Semantic Web: Research and Applications. Lecture Notes in Computer Science, Volume 4519/2007, pages 356–370, Springer-Verlag, June 2007.

II Osma Suominen, Eero Hyvönen, Kim Viljanen and Eija Hukka. HealthFinland – A National Semantic Publishing Network and Portal for Health Information. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, Volume 7, issue 4, pages 287–297, December 2009.

III Reetta Sinkkilä, Osma Suominen and Eero Hyvönen. Automatic Semantic Subject Indexing of Web Documents in Highly Inflected Languages. In *Proceedings of the 8th Extended Semantic Web Conference (ESWC2011)*, The Semantic Web: Research and Applications. Lecture Notes in Computer Science, Volume 6643/2011, pages 215–229, Springer-Verlag, June 2011.

IV Osma Suominen and Eero Hyvönen. Improving the Quality of SKOS Vocabularies with Skosify. In *Proceedings of the 18th International Conference on Knowledge Engineering and Knowledge Management (EKAW2012)*, Lecture Notes in Computer Science, Volume 7603/2012,

pages 383–397, Springer-Verlag, October 2012.

V Osma Suominen and Christian Mader. Assessing and Improving the Quality of SKOS Vocabularies. *Journal on Data Semantics*, forthcoming, June 2013.

In addition to these publications, this thesis references other work by the author [49, 44, 48, 118] to provide context and further information on the subjects discussed.

Author's Contribution

Publication I: "User-centric Faceted Search for Semantic Portals"

The author was the primary developer of the user-centric category elicitation process described in the paper and also performed all the experiments discussed. The procedure for mapping user-centric facets to underlying vocabularies was designed jointly with Kim Viljanen and Eero Hyvönen.

Publication II: "HealthFinland – A National Semantic Publishing Network and Portal for Health Information"

The author was the primary architect of the prototype system described in the article and designed the artifacts discussed, including the user interface. The distributed content creation framework, including the metadata schema and the requirements for content aggregation, was joint work with Eero Hyvönen and Kim Viljanen. The author performed all the experiments and wrote the majority of the text of the article. Eija Hukka participated as domain specialist on issues related to health information.

Publication III: "Automatic Semantic Subject Indexing of Web Documents in Highly Inflected Languages"

The author designed the research methodology and prepared the data sets for use in this study. Reetta Sinkkilä coordinated the work of human indexers and performed the automatic indexing experiments and Eero Hyvönen supervised the work. The analysis and reporting of results was performed jointly.

Publication IV: "Improving the Quality of SKOS Vocabularies with Skosify"

The author was responsible for this publication, and Eero Hyvönen supervised the work.

Publication V: "Assessing and Improving the Quality of SKOS Vocabularies"

The author was mainly responsible for the overall structure of the study, the evaluations performed using the PoolParty tool, and the aspects related to vocabulary quality improvement. Christian Mader was mainly responsible for defining the quality issues and for evaluating the vocabularies using the qSKOS tool. The selection of data sets and the reporting of results was joint work.

1. Introduction

1.1 Background and Research Environment

Semantic portals are information systems which collect information from several sources and combine them using semantic web technologies into a user interface that solves information needs of users [66, 113]. Lausen et al. [60] define semantic web portals as web portals, i.e., web sites that collect information for a group of users that have common interests and allow a community to share and exchange information, that are based on semantic web technologies. The use of semantic web technology [11], including technologies such as the Resource Description Framework (RDF) [17] and ontologies [30], allows the use of rich search functionality over structured data exposed on the web [90]. Early examples of semantic portals include SWED [99] and MuseumFinland [46].

Many semantic portals are based on the faceted search user interface paradigm (also known as faceted browsing, view-based search, dynamic hierarchies and guided navigation) [77], first developed in the HIBROWSE [93] and Flamenco [35] projects. A thorough review of the faceted search paradigm is given by Tunkelang [125] and another one (in Finnish) is given in the author's Master's Thesis [117]. The faceted search paradigm is especially powerful for the class of search tasks known as *exploratory search* [68], where the user's goal is to learn about and understand a particular topic, rather than looking up a specific item known in advance. Generic faceted search engines for semantic data such as BrowseRDF [86], /facet [40] and Longwell[1] have also been developed.

To date, many semantic portals have been published, especially as a result of academic research projects. These include the already mentioned

[1] `http://simile.mit.edu/longwell/`

SWED and MuseumFinland, as well as other domain-specific portals such as Promoottori [47], CS AKTive Space [110], mSpace Classical Music Explorer [105] and mSpace JSCentral [104], SW-Suomi.fi [111], MultimediaN E-Culture [106] and Orava [57]. Non-academic semantic portals include the BBC World Cup 2010 website[2], the Reegle energy portal[3] developed by the Semantic Web Company, and several portals developed by Mondeca[4].

The research presented in this dissertation summary has been performed at the Semantic Computing Research Group[5] as part of several research projects: the FinnONTO series of projects[6] (2003–2012), the SUBI project[7] (2009–2012), and the Linked Data Finland project[8] (2012–2014). These projects involved a series of semantic portal demonstration applications, which explored various aspects of semantic content production, quality assurance and user interface designs [48].

1.2 Objectives and Scope

This dissertation contains contributions in three areas related to semantic portals: faceted search user interface design, automatic subject indexing, and vocabulary quality. Each of these is addressed by a research question, presented in the following subsections.

1.2.1 User-centric Facets

While several semantic portals based on faceted search have been produced, their methodology for choosing or building suitable facets varies. The most common method [100], used in system such as SWED [99] and mSpace JSCentral [104], is to manually construct facets suited for the application domain. Some systems, including HIBROWSE [93] and Promoottori [47], use facets that are based directly on the underlying metadata and (possibly hierarchical) classifications.

However, a difficult situation arises when existing semantic metadata is repurposed for use in a faceted search interface; in this case, the controlled

[2] http://www.bbc.co.uk/blogs/bbcinternet/2010/07/bbc_world_cup_2010_dynamic_sem.html
[3] http://www.reegle.info
[4] http://www.mondeca.com/Clients
[5] http://www.seco.tkk.fi
[6] http://www.seco.tkk.fi/projects/finnonto/
[7] http://www.seco.tkk.fi/projects/subi/
[8] http://www.seco.tkk.fi/projects/ldf/

vocabularies referred to in the metadata may contain complex hierarchies that are unsuitable to be directly used in a faceted user interface [26], which would require the hierarchies not to be overly deep or wide [9]. The complexity is not necessarily a result of bad design, but arises because the vocabulary was not originally developed with browsing in mind. In addition, semantic search user interfaces have rarely been subjected to usability evaluations [7, 38], making it difficult to determine the actual usefulness of the faceted search paradigm for semantic portals. This leads into the third research question addressed in this dissertation:

3. **How can user-friendly search facets be constructed for semantic portals which are based on complex hierarchical vocabularies?**

To answer the first research question, related methods and findings from previous literature are first outlined in Section 2.1. In Section 3.1, a user-centric design method to solve this problem is presented, based on Publications I and II.

1.2.2 Automatic Subject Indexing

An important challenge in creating semantic portals is to obtain enough structured data. Many important data sources, such as document collections or event information databases, contain very little structured metadata. This makes it difficult to employ advanced search technologies and semantic web tools. *Information extraction* [18] is a set of methods, based on natural language processing techniques, which seek to obtain structured data from unstructured natural language texts, such as documents and event descriptions. Vocabulary-based automatic text categorization methods [108], also known as topic indexing, subject indexing, and term assignment, are in particular suited for creating simple structured metadata from unstructured documents and thus enabling better semantic search facilities.

To date, most automatic subject indexing tools have been developed either for English language texts or other (Romanic) languages with a relatively simple structure. In the FinnONTO context [45], however, many source documents are written in Finnish. Common natural language processing methods do not perform well with agglutinative and highly

inflected languages such as Finnish, Turkish, Estonian, Hungarian and Slavic languages [41, 64]. Therefore, new methods are needed that can handle highly inflected languages and produce good quality structured metadata from unstructured text. Preferably, the quality of automatic subject indexing should be as good as what human indexers produce, taking into account the varying levels of human performance and the relatively low agreement between any two humans indexing the same content. Human-competitive indexing quality has been previously achieved with English and some Romanic languages [71]. This leads into the second research question addressed in this dissertation:

2. **What is the quality of automatically assigned subjects for documents written in inflected languages compared with subjects assigned by human indexers?**

To answer the second research question, related methods and findings from previous literature are first outlined in Section 2.2. In Section 3.2, experimental results demonstrating that human-competitive quality of automatic subject indexing of Finnish language texts can be attained using a suitable combination of existing tools are presented, based on Publication III.

1.2.3 Vocabulary Quality

Finally, semantic portals based on faceted search, as well as many other kinds of semantic applications, depend on controlled vocabularies, which can be either thesauri [5], classifications or other types of controlled vocabularies such as lightweight ontologies [28]. Many recent semantic portals represent such vocabularies using the Simple Knowledge Organizing System[9] (SKOS) standard [8] for describing vocabularies by means of RDF structures. SKOS vocabularies can be used to integrate and interlink data from various sources by providing a common vocabulary. For example, many library classifications have been published as SKOS vocabularies, allowing various library catalogs using those classifications to be published as Linked Data and then easily integrated using RDF tools [14, 67, 116], enabling applications such as semantic information retrieval over multiple

[9] http://www.w3.org/2004/02/skos/

datasets [16], query expansion [123, 132], and recommendation.

However, the benefits of SKOS in data integration are only realizable if the SKOS vocabulary data is structurally valid and makes use of the SKOS entities in a meaningful way. To this end, the SKOS reference [8] defines a number of *integrity conditions* that can be used to detect inconsistencies in a SKOS vocabulary. In addition, *validation tools*, such as the PoolParty online SKOS Consistency Checker[10], are available for verifying that a SKOS vocabulary follows generally accepted best practices for controlled vocabularies which have not been codified in the SKOS reference. For example, many conceptual aspects of the desired structural qualities of thesauri and classifications, including the use of different types of hierarchical relationships, are discussed by Svenonius [119]. Some of these qualities, such as the overall structure and connectedness of the vocabulary, can be measured and evaluated algorithmically [53, 65].

Many SKOS vocabularies are currently published by automatically converting vocabularies from legacy formats into SKOS. Structural problems in the resulting SKOS files may be difficult to notice for vocabulary publishers, but may cause problems for users of the vocabularies [53, 65, 80]. This leads into the first research question addressed in this dissertation:

1. **How can the technical quality and validity of controlled vocabularies expressed in SKOS format be automatically measured and improved?**

To answer the third research question, related methods and findings from previous literature are first outlined in Section 2.3. In Section 3.3, a set of quality and validity criteria for SKOS vocabularies is first established, a representative set of publicly available SKOS vocabularies is evaluated against the criteria, and finally a method and a tool to correct many problems and deficiencies in the vocabularies is presented, based on Publications IV and V.

[10] `http://demo.semantic-web.at:8080/SkosServices/check`

1.3 Research Process and Dissertation Structure

1.3.1 Methodology

The main methodology used in this dissertation draws from the design science paradigm in information science [37, 91]. In the design science approach, a novel and innovative artifact to solve a relevant problem is 1) built using rigorous methods, 2) evaluated, 3) the design of the artifact is iterated when necessary, and finally 4) the results and lessons learned are effectively communicated. The created artifact can be, e.g., an information system, a computer program, or an algorithm that is useful for solving the identified problem, and its utility must be demonstrated by suitable design evaluation methods. The contribution of a design science research process is one or more of the following: 1) The design artifact itself; 2) Foundations, including design methods and algorithms; or 3) Methodologies, including novel evaluation methods and metrics. [37]

1.3.2 Demonstration Systems and Portals

Much of the research work contained in this dissertation was performed in the context of developing semantic portals and other demonstration applications:

The ONKI ontology library service[11] is a vocabulary publishing system for lightweight ontologies and SKOS vocabularies. It provides both human and machine access to the published vocabularies, including a browser user interface, Linked Data access, various Application Programming Interfaces (APIs), and vocabulary downloads [126]. The ONKI system has been online as a research prototype since the year 2005. A national production version of the system is currently being deployed.

The Sosiaaliportti[12] social workers' portal is a community web portal system developed and deployed by the Finnish National Institute for Health and Welfare[13]. In the FinnONTO research projects, semantic enhancements such as automatic subject indexing functionality were developed for the portal.

The HEALTHFINLAND portal (*TerveSuomi* in Finnish) is a semantic health information publishing system that aims to bridge the gap between

[11] http://www.seco.tkk.fi/tools/onki/
[12] http://www.sosiaaliportti.fi
[13] http://www.thl.fi

the health information needs of ordinary citizens and the expert organizations providing information about health on the web. The concept and ideas behind the portal was first formulated in two conference papers [44, 49]. The details of the first prototype system[14] were presented in the author's Master's thesis [117]. The system was since further developed into a production system[15] by the Finnish National Institute for Health and Welfare, and deployed in May 2009.

1.3.3 Datasets

The research described in Publications I and II addressing research question 1 was performed in the context of the HEALTHFINLAND demonstration system, using the document collection and controlled vocabularies of the portal as datasets.

The research in Publication III addressing research question 2 was performed with two document sets and vocabularies. Documents extracted from the Sosiaaliportti portal were described using concepts of the Finnish Ontology of Health and Welfare TERO, which is a lightweight ontology originally developed for the HEALTHFINLAND portal. Point of interest descriptions from Wikipedia were described using the Finnish Collaborative Holistic Ontology KOKO[16].

The research in Publications IV and V addressing research question 3 was performed with a representative selection of, in total, 33 publicly available SKOS vocabularies. Some of these vocabularies, including the Finnish General Thesaurus YSA[17], have been published via the ONKI ontology library system.

1.3.4 Dissertation Structure

The remainder of this dissertation summary is structured as follows. First, in Chapter 2, the state of the art in current research relevant for each research question is summarized. Second, in Chapter 3, the core research contributions of the publications contained in this dissertation are reviewed. Finally, in Chapter 4, the theoretical and practical implications of the studies are discussed, their reliability and validity assessed, and some recommendations for future research are presented.

[14] http://demo.seco.tkk.fi/tervesuomi/
[15] http://www.tervesuomi.fi
[16] http://www.seco.tkk.fi/ontologies/koko/
[17] http://www.nationallibrary.fi/libraries/thesauri/ysa.html

2. Theoretical Foundations

Research on semantic portals combines methods from multiple areas. In this section, related research is presented on three areas corresponding to the research questions: quality of controlled vocabularies, automatic subject indexing, and faceted search user interfaces.

2.1 User-centric Facets

Faceted classification schemes have a long tradition in the library science field, starting from the original idea formulated by S. R. Ranganathan in the 1930's [97]. The idea was further developed by Vickery into a practical system for implementing faceted classification schemes in specialist libraries [130].

2.1.1 Constructing Facets for Faceted Search

Constructing the facets for a faceted search application is an important part of the whole system design, because the facets affect the user interface, database design and data requirements. In earlier semantic portals based on the faceted browsing paradigm, the facets have been automatically created from the underlying taxonomies using projection rules (e.g., [131]).

A distinction can be made between systems that use pre-existing general purpose vocabularies and systems where the vocabularies are custom built with the intent to provide facets for the user interface. /facet[1] [40] is an example of the first approach, while the second group includes MuseumFinland [46] and SWED[2]. The problems of matching the hierarchical structure of the vocabulary with user needs and expectations only become apparent in the first case, as the point of view of the original vocabulary may differ

[1] `http://slashfacet.semanticweb.org`
[2] `http://www.swed.org.uk`

a lot from the end-users' mental models of the information space. In /facet, the automated facet generation sometimes results in a user interface that is hard to use [40].

Another approach for creating a navigational hierarchy based on an existing controlled vocabulary is presented by Stoica and Hearst [115]. Their system uses the WordNet lexical database as a basis for creating a hierarchical classification which can then be used in faceted browsing. The Castanet algorithm simplifies the WordNet IS-A hierarchy by eliminating branches that aren't represented in the document collection as well as unnecessary levels of the hierarchy. The resulting taxonomies can be used either as-is or after some manual adjustments. However, the relationship of Stoica and Hearst's work with metadata that references existing controlled vocabularies is weak: WordNet is only used as a basis for creating the navigational hierarchies, and the document metadata is later assumed to reference the newly created taxonomy directly.

2.1.2 User-centered Design Methods

User-centered design is a methodology which seeks to understand actual needs of users and to design systems, including their functionality and user interfaces, accordingly [56, 59, 61]. A special field of user interface design relevant to web navigation and search interface design is *information architecture* [74, 100], which comprises a set of methods and design practices to help organize large content spaces in a way that makes them easy to navigate.

Card sorting [70, 101, 112] is an important information architecture design method that helps elicitate users' mental models of information spaces. In card sorting experiments, current or future users of a system organize decks of cards (either physical or virtual) into groups that make sense to them. The sort results are then analyzed, often with a spreadsheet template [58, 112]. The results of the analysis are used to construct information spaces, for example navigation structures of intranets [10, 84, 124] or other large web sites [100, 112].

Card sorting has also been used in the construction of ontologies as a means of knowledge elicitation [101, 109]. While card sorting is usually performed manually outside the ontology engineering process, a computerized card sorting plugin has been developed for the Protégé[3] ontology

[3] `http://protege.stanford.edu`

editor [134]. However, the focus of this work is on the ontology creation process itself; there is no direct intent of using the resulting ontology in a search-oriented user interface.

2.2 Automatic Subject Indexing

Subject indexing is the process of describing the topic or main subject matter of documents using terms or concepts from a pre-defined controlled vocabulary such as a thesaurus. It is traditionally performed by humans, for example as part of the process of cataloging documents in a library. The assigned subjects can be later used to retrieve the documents based on their topics. The principles for subject-based document retrieval were developed in the late 19th and early 20th century by library science pioneers including Melvil Dewey [22], Charles A. Cutter [21], S. R. Ranganathan [97], and Paul Otlet [87].

In *automatic subject indexing*, the task of assigning subjects to documents is given to an algorithm, which has been practised since the early 1960's [69]. It is part of *information extraction*, which is the practice of isolating structured information from unstructured natural language text [18]. A closely related field is *named entity recognition*, which is concerned with recognizing textual references to real world entities such as person, organization and location names, as well as numeric expressions including time, date, money and percent expressions [79]. Tools such as ANNIE, a part of the GATE[4] natural language processing toolkit [20], and KIM [52, 94], which performs information extraction using semantic web technologies and is also based on GATE, can be used to perform named entity recognition and other information extraction tasks.

Many automatic subject indexing tools exist for various languages and domains [108]. For example, many systems have been developed for assigning subjects from the Medical Subject Headings (MeSH) vocabulary to biomedical documents [122]. General purpose automatic subject indexing tools, which can be used with any controlled vocabulary in any domain, include Maui [71]; its predecessors KEA [135] and KEA++ [72]; the HIVE system [29], which incorporates the KEA algorithm; and the PoolParty Extractor system[5]. These tools can also perform topic indexing without the support of a controlled vocabulary, known as *keyphrase extraction*.

[4] `http://gate.ac.uk`

[5] `http://www.poolparty.biz/portfolio-item/poolparty-extractor/`

On a high level, automatic subject indexing consists of two phases: first, performing linguistic analysis for matching document words or n-grams with meanings expressed as terms in a controlled vocabulary (*semantic tagging*), and second, determining which of the matched vocabulary terms best describe the document (*topic ranking*).

2.2.1 Semantic Tagging

Semantic tagging is the matching of words to meanings and part of linguistic analysis. Linguistic analysis for the purpose of annotation consists of five steps: morphological analysis, part-of-speech tagging, chunking, dependency structure analysis and semantic tagging [15]. In languages such as English, Spanish and French, a simplified form of semantic tagging can be performed by using a rule-based stemming algorithm to normalize both document words and vocabulary terms [71]. This allows, e.g., singular words to be matched with plural terms in the vocabulary. Well-known stemming algorithms include the Lovins [63] and Porter [95] stemmers, the Snowball stemmer[6], and Koskenniemi's two-level model for morphological analysis [55].

Inflected languages such as Finnish, Turkish, Arabic and Hungarian typically express meanings through morphological affixation. In highly inflected languages plural and possessive relations, grammatical cases, and verb tenses and aspects, which in English would be expressed with syntactic structures, are characteristically represented with case endings [64, 85, 121]. Compound words are also typical in inflected languages. Rule-based stemming does not work particularly well for analyzing inflected languages [6, 54]: for example, a semantic tagger for the Finnish language developed in the Benedict project used a sophisticated morphological analysis and lemmatisation tool as well as rules for handling compound words in order to attain high precision [64]. However, in probabilistic information retrieval of Finnish documents, a stemmer can perform as well as a lemmatization algorithm [51].

2.2.2 Topic Ranking

The TF×IDF method provides a widely used baseline for ranking topics [102]. In topic ranking, machine learning methods have surpassed rule-based methods for determining the important topics of a document [108].

[6] `http://snowball.tartarus.org/texts/introduction.html`

KEA, KEA++ and Maui have improved on TF×IDF ranking by additionally using various heuristics and machine learning.

KEA has been ported to support other languages. A Turkish adaptation of KEA was used to extract keyphrases without using a controlled vocabulary [89]. A KEA-like approach for keyphrase extraction of Arabic documents has also been found to perform well when part-of-speech analysis was incorporated into the candidate selection phase [25].

In tests on English, French and Spanish documents, Maui has been found to assign subjects of comparable quality of those of humans [71]. In these tests, a stemming algorithm was used to aid basic semantic tagging.

Other subject indexing tools for inflected languages include the Poka information extraction tool for Finnish [127], which has been used in the Opas system to assign concepts from the Finnish General Upper Ontology to question-answer pairs [129]. The Leiki platform is a commercial tool that analyzes Finnish text and attempts to determine its important concepts using a proprietary ontology-like classification system [92]. It is used by some Finnish news websites for generating links to related content. However, neither tool has been evaluated in academic literature.

2.3 Vocabulary Quality

Controlled vocabularies such as thesauri, classifications, term lists, and lightweight ontologies, were first developed in libraries, for classifying books and other documents in library collections. Early controlled vocabularies include the Dewey Decimal Classification[7], first published in 1876 by Melvil Dewey and still by far the most popular method of organizing library collections; the Library of Congress Classification[8], also dating from the late 19th century; and the Library of Congress Subject Headings[9] (LCSH), a thesaurus for maintaining bibliographic records first published in 1909. Controlled vocabularies used outside the library sector include the Art and Architecture Thesaurus[10], used to index museum collections; the Medical Subject Headings[11] vocabulary, used for indexing biomedical

[7] http://www.oclc.org/dewey/
[8] http://www.loc.gov/catdir/cpso/lcc.html
[9] http://id.loc.gov/authorities/subjects.html
[10] http://www.getty.edu/research/tools/vocabularies/aat/
[11] http://www.nlm.nih.gov/mesh/

documents in online catalogs such as PubMed[12]; and the AGROVOC[13] thesaurus published by the Food and Agriculture Organization of the United Nations, used for information management in many databases related to agriculture, forestry, fisheries, environment and related domains.

Controlled vocabularies are useful tools in organizing large-scale web information systems [100]. Thus, they are used in many kinds of semantic applications, including semantic search systems [38], annotation tools [52], and semantic portals [60]. Early semantic web applications often used the RDF Schema [33] and/or the Web Ontology Language (OWL) [107] to express controlled vocabularies, such as in KIM [52], MuseumFinland [46], and the MultimediaN E-Culture demonstrator [106].

Starting in 2004–2005, SKOS [8] has emerged as a practical language for expressing controlled vocabularies as RDF data, and has been used in many systems including SWED [99], Semport [19], Sowiport [13] and HEALTHFINLAND to express controlled vocabularies. Hundreds of controlled vocabularies expressed using SKOS have been made available on the Web of Data [4].

2.3.1 Quality of SKOS Vocabularies

An early guide for creating SKOS vocabularies by Miles et al. [73] already stressed the importance of error checking and validation, but the validation is only performed on the RDF syntax level. Van Assem's description of a method for converting existing thesauri to SKOS [128] notes the difficulty of SKOS validation, which has since been addressed by later revisions of the SKOS specification and the development of validation tools.

The SKOS reference specifies in total six *integrity conditions*, which must be fulfilled for the vocabulary to be considered valid [8]. Many of these conditions are based on earlier standards for structuring controlled vocabularies and thesauri, including ISO 2788 [1] and the British standard BS8723 Part 2 [2]. These conditions may be considered a minimum set of validation and/or quality criteria for SKOS vocabularies; there are also many vocabulary-related best practices which go beyond the integrity conditions codified in SKOS.

Kless and Milton [53] provide an overview about intrinsic abstract measurement constructs for thesaurus evaluation. Nagy et al. have explored the various structural requirements of SKOS vocabularies in different

[12] `http://pubmed.gov`
[13] `http://www.fao.org/agrovoc/`

application scenarios [80]. Mader et al. have developed a more extensive set of quality criteria for SKOS vocabularies in the qSKOS project[14] [65].

2.3.2 Evaluating SKOS Vocabularies

The *PoolParty online SKOS Consistency Checker* (hereafter known as the PoolParty checker) is an online validation tool performs many checks on SKOS vocabularies, including the SKOS integrity conditions. It has originally been developed to determine if the vocabulary can be imported into the *PoolParty thesaurus editor* [103]. The W3C used to host a similar online SKOS validation service, but it was not kept up to date with the evolution of SKOS, and is no longer available. However, implementations vary, particularly in the level of support for RDFS and OWL reasoning, SKOS inference rules, and the extent to which they implement the informally specified SKOS integrity conditions. Thus, the results of these checks cannot always be directly compared.

Abdul Manaf et al. [4] have surveyed the landscape of SKOS vocabularies available on the Web and analyzed their high level structural properties, such as the number of hierarchy levels and in- and outgoing links to other concepts. The same authors have also identified three types of common problems (*slips*) in SKOS vocabularies as well as possible ways to correct them (*patches*) [3]. They can be found by OWL reasoning and are partly based on the axioms defined in the SKOS reference ontology. However, the number of proposed slips and corresponding patches is quite small and mostly concerned with making the SKOS vocabularies processable using an OWL reasoner, not with the quality of the intellectual content of the vocabulary.

The authors of the SKOS version of the STW Thesaurus of Economics describe the use of SPARQL queries to find inconsistencies in SKOS vocabularies [81]. However, they do not describe the consistency checks they used in detail.

2.3.3 Quality of Linked Data Sets

More general validation services for RDF and Linked Data have also been developed. The *W3C RDF Validation Service*[15] can be used to verify the syntax of RDF documents. The *Vapour* [12] system is intended to spot

[14] `https://github.com/cmader/qSKOS`
[15] `http://www.w3.org/RDF/Validator/`

problems with HTTP content negotiation in published RDF and Linked Data. The *RDF:Alerts* [42] system is another online validation tool that can be used to spot syntax errors, inconsistencies, incomplete data, misuse of classes and properties, and other kinds of problems in Linked Data. For OWL datasets, the *Pellet ICV* reasoner has a validation mode in which re-interprets OWL axioms with integrity constraint semantics and can thus be used to find inconsistencies in RDF data involving OWL axioms.

The SPARQL Inferencing Notation[16] (SPIN) is a SPARQL-based language which can be used to specify integrity constraints for RDF data [27]. The *TopBraid Composer*[17] suite is one tool supporting SPIN-based validation, and it includes a SPIN ruleset that implements testing of the SKOS integrity conditions.

A recent and thorough survey of general RDF and Linked Data validation tools is given by Hogan et al. [42] identifying four categories of common errors and shortcomings in RDF documents. Also, Heath et al. [36] summarize best practices for publishing data on the Web. The *Pedantic Web Group*[18] is an online community of practitioners who help to correct errors in the publication of RDF data. However, to the author's knowledge, none of these tools and approaches have any specific support for SKOS vocabularies.

2.3.4 Ontology Evaluation, Repair, and Improvement

Ontology evaluation, i.e., measuring the quality of an ontology, was introduced when ontologies were first put to use in information systems [30]. Competency questions for evaluating ontologies were proposed by Grüninger and Fox [31]. Principles for creating good quality ontologies were further developed by researchers developing ontologies for use in the biomedical field [98, 114]. The OntoClean methodology introduced a set of guidelines for validating ontologies to expose inappropriate or inconsistent modeling choices [32]. An exhaustive modern discussion of ontology evaluation has been provided by Vrandecic [133].

Repairing problematic constructs in OWL ontologies has been extensively discussed by Kalyanpur [50]. Ovchinnikova et al. propose a method for solving inconsistencies in ontology design by rewriting problematic axioms [88]. Horridge et al. present methods for explaining inconsistencies in OWL

[16] `http://spinrdf.org`
[17] `http://www.topquadrant.com/products/TB_Composer.html`
[18] `http://pedantic-web.org`

ontologies [43]. The OOPS! pitfall scanner is an OWL ontology evaluation tool that provides the user with guidelines about how to solve the issues it has found [96].

However, these OWL-related methods are only partially relevant to SKOS vocabularies, because not all of the SKOS integrity conditions and other quality measures can be expressed using OWL axioms[19]. To the author's knowledge, automatic correction methods intended specifically for SKOS vocabulary constructs have not been proposed earlier.

[19] In particular, neither OWL nor OWL 2 include any means to express the integrity condition S14: *"A resource has no more than one value of* `skos:prefLabel` *per language tag."*

3. Research Contributions

The current state of the art in the three areas described in the previous chapter leaves some questions unanswered, particularly for scenarios where semantic portals are built for layman end users for whom complex vocabulary hierarchies may be problematic, with incomplete metadata about documents in inflected languages, using controlled vocabularies with possible quality issues. This chapter presents solutions to those challenges based on the publications included in this dissertation.

3.1 User-centric Facets

Publications I and II together address research question 3: *How can user-friendly search facets be constructed for semantic portals which are based on complex hierarchical vocabularies?*

The main results of the study described in the two publications are a method for creating user-centric facets for information systems based on metadata that references controlled vocabularies, and a method for mapping these facets into underlying vocabularies in order to create a functional faceted search user interface. Publication II gives an overview of the HEALTHFINLAND prototype system, the context in which the method was developed, and describes the evaluations performed during and after implementing these methods in the HEALTHFINLAND portal prototype. Publication I describes the methods and their specific evaluation procedures in more detail.

3.1.1 Creating User-centric Facets

The process for creating user-centric facets, presented in Publication I, is outlined in Table 3.1. The process consists of six sequential steps. The last two steps can be iterated several times. The process is an adaptation of

Table 3.1. Process for creating user-centric facets from controlled vocabularies and document metadata.

Step	Title	Description
1	Select card contents	Select concepts from vocabulary, based on existing metadata.
2	Perform card sort	Recruit around 10 representative users of the system and ask them to sort the cards into groups that make sense to them.
3	Analyze sort results	Cluster the different user categorizations into standard categories. Analyze using spreadsheet template.
4	Design initial categories	Choose top-level categories based on card sort. Fill in lower levels with concepts drawn from the vocabulary.
5	Evaluate categories	Possible evaluation methods include closed card sorting, expert review and usability tests on system prototypes.
6	Finalize categories	Remedy the problems found in evaluation. Repeat steps 5–6 as necessary.

a typical card sorting approach, where the cards are typically based on documents or important topics drawn from the content of a web site [70, 112]. The novelty is the use of concepts found in the controlled vocabularies referenced in content metadata as source material for the cards, which allows the final categories to be later mapped to concepts drawn from the original vocabulary.

3.1.2 Mapping User-centric Facets to Vocabulary Concepts

In order to implement faceted search over metadata using user-centric facets, their relationship to the underlying vocabularies must be explicitly represented. In Publication I, our mapping solution based on the SKOS Core [8] and SKOS Mapping[1] vocabularies is presented. The explicit representation of the mappings using RDF vocabularies enables the implementation of a faceted search engine.

An example illustrating the use of mappings to represent the relationship between facets and underlying vocabularies is shown in Figure 3.1. In the example, the facet category *Weight control* is mapped directly to the MeSH concept *Body Weight*, as well as indirectly via the subcategory *Losing weight* to MeSH concepts *Weight Loss* and *Energy Intake*[2]. Likewise, the facet category *Nutrition & Food* is mapped to the MeSH concept *Energy In-*

[1]The mapping vocabulary was merged into SKOS Core after the research was conducted.

[2]The `narrowMatch` mapping between *Losing weight* and the MeSH concept *Energy Intake* is not strictly correct in this example, taken from Publication I, as energy intake may also be considered in other contexts than weight loss. The example is based on actual mappings generated using the card sorting technique.

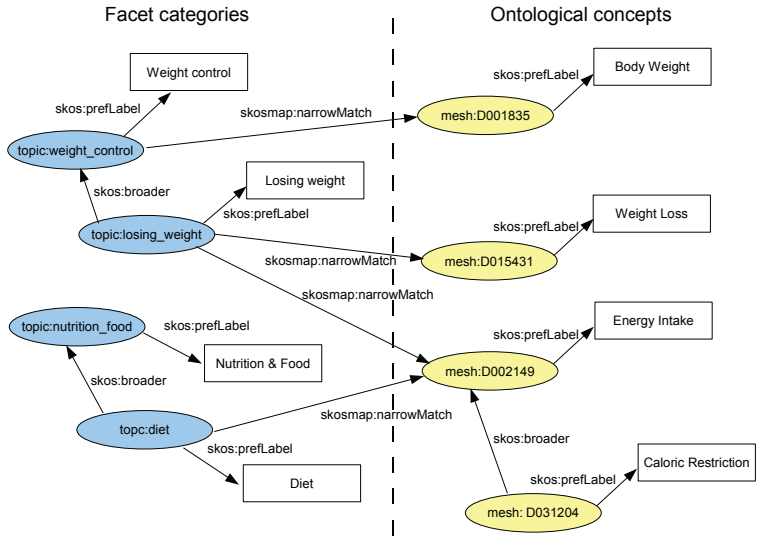


Figure 3.1. Examples of mappings between facet and vocabulary concepts. The URI prefix `topic` refers to the *Topic* facet and `mesh` to the indexing vocabulary MeSH.

take through the subcategory *Diet*. In the user interface, any content items described using the MeSH concept *Caloric Restriction*, whose `broader` concept is *Energy Intake*, would be visible in all the facet categories shown in the example, while content described using the MeSH concept *Body Weight* would only be shown in the facet category *Weight control*. The novelty in this method is the use of an intermediate facet layer, which allows for a user-centric view into structured metadata while preserving the full expressibility of the underlying conceptual representation, instead of confronting the user with complex hierarchical structures.

3.1.3 User-centric Faceted Search Engine

Publication I presents the initial prototype of the faceted search engine of the HEALTHFINLAND portal, while Publication II presents its evolution into the final HEALTHFINLAND production system (Figure 3.2).

3.1.4 Evaluation and Lessons Learned

The method for creating user-centric facets was evaluated by doing a limited closed card sort experiment on the initial categorization, as well as an expert review, detailed in Publication I. The prototype was evaluated through multiple user studies, including a series of usability tests discussed in Publication II.

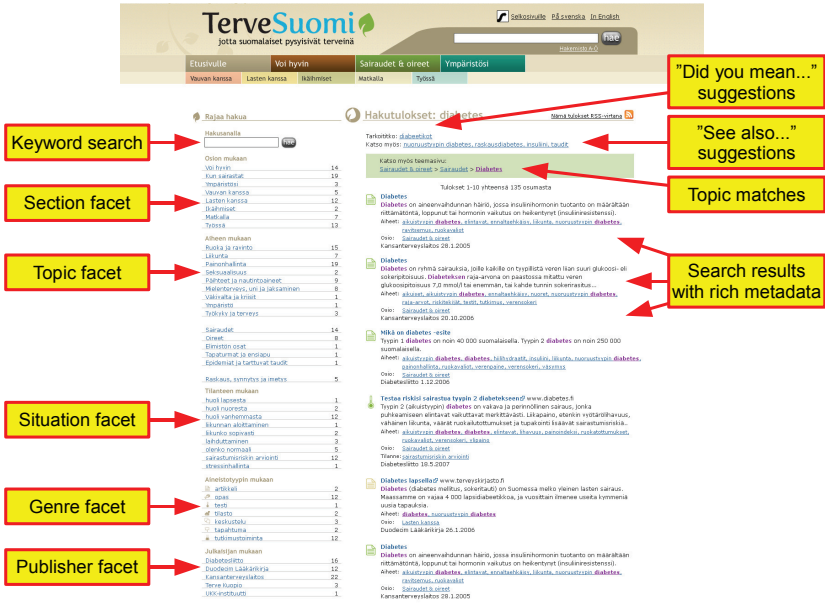


Figure 3.2. Faceted search interface of the HEALTHFINLAND production system.

The main lesson learned from the evaluations was that the user-centric categories produced by the method were intuitive and useful on the top levels, but users sometimes struggled with the lower levels because they were constructed directly from the underlying vocabularies.

Another finding was that splitting up a single large vocabulary into separate facets, as was done in the HEALTHFINLAND prototype system, will cause problems with the search interface because content items are unlikely to be consistently labeled with specific values for all facets. The problem can be addressed to some degree with default values, but still, some of the power of faceted search will be lost due to missing information. Due to this issue, the facets for the HEALTHFINLAND production system were redesigned so that vocabularies were not split and the metadata schema contained explicit fields for each facet. In the new portal, both the browsing and search user interfaces were based on the categorization of documents by Section (audience), Topic, Situation, Genre and Publisher. Each of these categorizations for a document were directly described as a separate field in the metadata used by the production system, instead of using a more general subject property as in the prototype system.

Table 3.2. Stemming and lemmatization strategy results

	Precision	Recall	F-measure
SOS-60, FDG lemmatizer	40.0	37.1	38.5
SOS-60, Omorfi lemmatizer	40.0	35.9	37.8
SOS-60, Snowball stemmer	35.7	32.2	33.8
French Agrovoc [71]	34.5	31.8	33.1
Spanish Agrovoc [71]	24.7	26.9	25.7

3.2 Automatic Subject Indexing

Publication III addresses research question 2: *What is the quality of automatically assigned subjects for documents written in inflected languages compared with subjects assigned by human indexers?*

The results of the study consist of three parts, each testing the state-of-the-art Maui framework [71] for automatic subject indexing using Finnish-language documents: 1) a comparative evaluation of three stemming and lemmatization algorithms; 2) an evaluation of the consistency between human indexers and the Maui algorithm; and 3) an experiment to verify the domain independence of Maui indexing.

3.2.1 Stemming and Lemmatization Strategy

The first experiment tested the suitability of the Maui tool for the Finnish language with alternating stemmers. The results of experiment, summarized in Table 3.2, demonstrated that of the three stemmers tested, both Omorfi [62] and FDG [120] can be used for lemmatization and both will give results that are as good or better than those obtained using comparable tools for other languages. The best lemmatisation strategy was FDG, but Omorfi was not far behind. The simple rule-based Snowball stemming algorithm was used as a baseline. The precision, recall and F-measure values attained were higher than the measurements attained for French and Spanish documents in the original Maui evaluation [71], which have been included in Table 3.2 for comparison. While the values taken in isolation may appear rather low (the theoretical maximum for each being 100), the results compare well with measurements of consistency between human indexers, which was the subject of the second experiment.

Table 3.3. Consistency of human indexers 1–6 compared to Maui

	1	2	3	4	5	6	Average	Maui
1		25	29	28	27	28	27.4	21.5
2	25		31	30	36	37	31.8	29.9
3	29	31		40	42	39	36.2	27.2
4	28	30	40		38	35	34.2	36.3
5	27	36	42	38		40	36.6	25.3
6	28	37	39	35	40		35.8	27.2
							33.7	27.9

3.2.2 Inter-indexer Consistency

The second experiment found that the Maui topic indexing algorithm is 27.9% consistent with human indexers, while the consistency between any two human indexers was 33.7% on average. With a suitable lemmatization tool, the performance of Maui in terms of agreement with human indexers is almost on the same level as that of the human indexers themselves.

3.2.3 Domain Independence

The results of our third experiment, using point of interest descriptions and a general lightweight ontology, suggest that when a suitable lemmatizer is used the algorithm also works well with Finnish text of different domains.

3.3 Vocabulary Quality

Publications IV and V together address research question 1: *How can the technical quality and validity of controlled vocabularies expressed in SKOS format be automatically measured and improved?*

Both studies consist of three main parts: 1) a synthesis of quality and validity criteria for SKOS vocabularies; 2) an analysis of published SKOS vocabularies according to the criteria and 3) a methodology and a tool for correcting different kinds of common problems in SKOS vocabularies.

The study described in Publication IV used 14 vocabularies, 11 quality criteria, and two tools for vocabulary analysis. Publication V describes an expanded follow-up study, with 24 vocabularies, 26 quality criteria, and three analysis tools. There is some overlap between the vocabulary data sets, so in combination, the studies cover 33 vocabularies.

3.3.1 Validation Criteria

Publication IV presents a synthesis of validity and quality criteria based on earlier sources and tools, including the PoolParty checker, the qSKOS framework for SKOS vocabulary quality analysis, and the integrity conditions defined by the SKOS specification. The result is a list of 11 quality criteria, of which 9 are based on the PoolParty checker.

Publication V defines a more comprehensive set of quality criteria, mainly based on the work by Mader et al. in developing the qSKOS vocabulary quality analysis methodology and toolkit. The resulting 26 quality criteria are summarized in Table 3.4. More formal definitions of the criteria are given in Publication V.

Table 3.4. Criteria for assessing the quality and validity of SKOS vocabularies.

Criterion	Description
Omitted or Invalid Language Tags	Natural language labels specified without a valid, explicit language tag.
Incomplete Language Coverage	Concepts lacking labels in some of the languages that exist in the vocabulary.
Undocumented Concepts	Concepts without any SKOS documentation properties.
Overlapping Labels	Multiple concepts with the same label.
Missing Labels	Concepts and other vocabulary constructs not given a human-readable label.
Inconsistent Preferred Labels	Concepts with multiple preferred labels in the same language, violating the SKOS integrity condition S14.
Disjoint Labels Violation	The number of SKOS labels that violate the labeling disjointness axiom S13 defined by the SKOS specification.
Extra Whitespace in Labels	The number of literal terms which contain extra surrounding whitespace.
Orphan Concepts	Concepts without any associative or hierarchical relationships.
Disconnected Concept Clusters	Separate concept clusters disconnected with the main vocabulary.
Cyclic Hierarchical Relations	The number of cycles in the vocabulary hierarchy.
Valueless Associative Relations	Sibling concepts having an associative relationship, if that relationship is only justified by the concepts being siblings.
Solely Transitively Related Concepts	Concepts linked by SKOS `broaderTransitive` and/or `narrowerTransitive` relationships, without being linked by (chains of) `broader` and/or `narrower` relationships that would justify the transitive relationships.
Omitted Top Concepts	Concept schemes lacking any explicitly identified top-level concepts.
Unmarked Top Concepts	Top-level concepts in the vocabulary that are not identified explicitly.
Top Concepts Having Broader Concepts	Concepts marked as top concepts that are not actually the topmost concepts in the hierarchy.
Unidirectionally Related Concepts	Concepts having only a one-way relationship, when SKOS defines an inverse relationship that should also exist in the vocabulary.
Relation Clashes	SKOS semantic relations defined as disjoint by the SKOS integrity condition S27 that are incorrectly used together.
Mapping Clashes	SKOS mapping properties defined as disjoint by the SKOS integrity condition S46 that are incorrectly used together.
Disjoint Classes Violation	SKOS constructs that violate the class disjointness axioms S9 and S37 defined by the SKOS specification.
Missing In-links	Concepts that are not linked to from any public repositories of semantic data.
Missing Out-links	Concepts that do not link to any external data sets.
Broken Links	Concepts whose URI is not dereferenceable on the Web of Data.
Undefined SKOS Resources	References to constructs in the SKOS namespace that are not actually defined by the SKOS specification.
HTTP URI Scheme Violation	Non-HTTP URIs used in the vocabulary.
Invalid URIs	Whether the URIs used in identifying concepts and other SKOS vocabulary concepts follow specifications and best practices for choosing URIs.

Table 3.5. Results of validating the vocabularies using the PoolParty checker before and after performing corrections with Skosify.

	Publication	Valid URIs	Missing Language Tags	Missing Labels	Loose Concepts	Disjoint OWL Classes	Consistent Use of Labels	Consistent Use of Mapping Properties	Consistent Use of Semantic Relations
EARTh	V	pass	pass	**fail**	2687→pass	pass	**fail**→pass	pass	**fail**→pass
GBA	IV	pass	pass	pass	pass	pass	pass	pass	pass
GEMET	IV,V	pass	3→pass	pass	109→pass	pass	**fail**→pass	pass	**fail**→pass
Geonames	V	pass	pass	pass→**fail**	pass	pass	**fail**→pass	pass	pass
IPSV	V	pass	pass	**fail**	pass	pass	**fail**→pass	pass	**fail**→pass
IPTC	V	pass	pass	pass	pass	pass	pass	pass	pass
IAUT93	IV	pass	358→pass	**fail**	1060→pass	pass	**fail**	pass	**fail**→pass
IVOAThes.	IV	pass	2890→pass	pass	926→pass	pass	pass	pass	**fail**→pass
LVAk	V	pass	13411→pass	pass	69→pass	pass	pass	pass	**fail**→pass
MeSH2006	IV	pass	pass	pass	189→pass	pass	pass	pass	**fail**→pass
NASA	IV	pass	88→pass	pass	1→pass	pass	pass	pass	pass
NYTL	IV,V	pass	pass	pass	1920→pass	pass	**fail**	pass	pass
NYTP	V	pass	pass	pass	4979→pass	pass	pass	pass	pass
NYTS	IV	pass	pass	pass	498→pass	pass	pass	pass	pass
ODT	V	pass	pass	pass	pass	pass	**fail**→pass	pass	pass
Plant	V	pass	pass	pass	pass	pass	pass	pass	pass
PXV	V	pass	1684→pass	**fail**	7→pass	pass	**fail**→pass	pass	**fail**→pass
Reegle	V	pass	pass	pass	2→pass	pass	**fail**	**fail**	**fail**→pass
SNOMED	V	pass	102599→pass	**fail**	pass	pass	**fail**→pass	pass	**fail**→pass
ScOT	IV	pass	pass	pass	pass	pass	**fail**	pass	**fail**→pass
SSW	V	pass	pass	pass	9→pass	pass	**fail**→pass	pass	**fail**→pass
STW	IV,V	pass	2→pass	**fail**	pass	pass	**fail**→pass	pass	**fail**→pass
UMBEL	V	pass	25794→pass	pass→**fail**	pass	pass	**fail**→pass	pass	pass
UNESCO	V	pass	pass	pass	pass	pass	pass	pass	pass
YSA	IV	pass	pass	**fail**	8614→pass	**fail**→pass	pass	pass	**fail**→pass

3.3.2 Validity of SKOS Vocabularies

33 published SKOS vocabularies were altogether analyzed in the studies using the PoolParty checker, the qSKOS quality analysis toolkit, and the Skosify tool. Many of the vocabularies were found to contain structural problems, including violations of the integrity conditions defined by the SKOS specification. The results of validating 25 of the vocabularies using the PoolParty checker tool are summarized in Table 3.5. Vocabularies that were too large for the PoolParty checker were omitted from the table.

In Publication V, 24 vocabularies were analyzed with the qSKOS tool. The results of the analysis before and after performing corrections with the Skosify tool are summarized in Tables 3.6, 3.7, and 3.8. The vocabularies have been sorted alphabetically, as in Table 3.5, but otherwise the data is the same as presented in Publication V. Not all the quality criteria listed in Table 3.4 are included in these results, because some of the checks were not implemented in the qSKOS tool.

In both studies, around three quarters of the examined vocabularies were found to violate one or more of the SKOS integrity conditions. In particular, both studies show that the SKOS integrity condition S27, which specifies that the `related` relationship is disjoint with the `broaderTransitive` relationship, is violated by the majority of the vocabularies that were examined.

In the publications, an amendment to the SKOS specification was suggested that would specify that `related` is disjoint with `broader`, not the transitive variant. This would prevent the more benign cases of current S27 integrity condition violations from being considered errors, thus increasing the availability of structurally valid SKOS vocabularies on the Web of Data.

We also found that performing full RDFS and OWL inference is important for finding some quality issues. The three vocabulary evaluation tools we used had varying levels of support for inference, which sometimes caused differing results. For example, in Publication V, some inconsistent labels in the New York Times Locations (NYTL) vocabulary were only found by the PoolParty checker, because it is the only tool that performs `owl:sameAs` inference.

3.3.3 Correcting Problems

The publications present a methodology and a tool, Skosify, for correcting structural problems in SKOS vocabularies. The tool was able to correct the great majority of structural problems in the vocabularies identified by the PoolParty tool, as shown in Table 3.5. Eight of the quality issues identified by the qSKOS tool were targeted by the correction heuristics implemented in Skosify. For these quality issues, Skosify was similarly able to correct the great majority of issues, as shown in Tables 3.6 and 3.7.

Table 3.6. Validation and correction results using the qSKOS quality analysis toolkit, part 1: *Labeling and Documentation Issues*. The figure for *Extra Whitespace in Labels* was determined using the Skosify tool.

	Omitted or Invalid Language Tags	Incomplete Language Coverage	Undocumented Concepts	Overlapping Labels	Inconsistent Preferred Labels	Disjoint Labels Violation	Extra Whitespace in Labels
AGROVOC	0	32060	29820	2666→2683	0	2424→0	2166
DBpedia	0	0	865902	765	0	0	0
DDC	0	158161	251977	40729	1→0	0	416
EARTh	10→0	313	7840	2100→2103	0	69→0	310
Eurovoc	219	6370	5341	62	0	0	2
GEMET	4→0	894	1	3638	0	3→0	12
GeoNames	0	43	60	162	1→0	0	0
GTAA	0	0	96850	11894	0	0	0
IPSV	0	0	4551	0	0	21→0	0
IPTC	0	0	933	1	0	0	0
LCSH	100316→0	0	308607	7766	669→0	206→0	0
LVAk	13411→0	0	13411	13	0	0	0
NYTL	0	0	1862	0	0	0	0
NYTP	0	0	4094	0	0	0	6
ODT	3→0	16	35	2	0	1→0	0
Plant	1→0	0	220	54	0	0	0
PXV	1578→0	0	1492	7	0	4→0	2
RAMEAU	116343→0	140860→172469	70358	5539→5905	0	33066→0	7940
Reegle	3→0	1450	3	22	0	3→0	52
SNOMED	102600→0	0	102614	229	0	202→0	0
SSW	4→0	1143	1328	39	0	16→0	6
STW	47→45	25050	5290	10123	214→0	0	0
UMBEL	25793→0	0	2848	5207→5226	2→0	1→0	522
UNESCO	0	0	2509	227→279	0	0	1524

Table 3.7. Validation and correction results using the qSKOS quality analysis toolkit, part 2: *Structural Issues*.

	Orphan Concepts	Disconnected Concept Clusters	Cyclic Hierarchical Relations	Valueless Associative Relations	Solely Transitively Related Concepts	Omitted Top Concepts	Top Concepts Having Broader Concepts	Unidirectionally Related Concepts	Relation Clashes	Mapping Clashes
AGROVOC	0	234	0	281	0	0	0	20672→0	1→0	0
DBpedia	103877→103880	1174→1171	1133	9021→6352	0	0	0	1713339→0	10219→0	0
DDC	97294	2087	0	0	0	30→5	1812	4761→0	0	0
EARTh	2288	354	0	1124	0	0	0	12091→0	61→0	0
Eurovoc	7	4	0	6→5	0	1→0	0	14289→0	0	0
GEMET	0	5	0	31	0	1→0	0	9657→0	2→0	0
GeoNames	680	0	0	0	0	9→0	0	0	0	0
GTAA	162000	621	0	9448→9414	0	9→0	0	18804→0	37→0	0
IPSV	0	1	0	253	0	0	0	25→0	5→0	0
IPTC	0	10	0	0	1113→0	0	0	2241→0	0	0
LCSH	173149	22343	0	0	0	1→0	0	96533→0	0	0
LVAk	21	11	5→0	5	0	0	0	16344→0	1→0	0
NYTL	1920	0	0	0	0	1→0	0	0	0	0
NYTP	4979	0	0	0	0	1→0	0	0	0	0
ODT	4	7	0	7→6	0	0	2	126→0	0	0
Plant	0	22	0	3463	0	0	44	3246→0	0	0
PXV	2	10	0	0	0	0	1	2725→0	2→0	0
RAMEAU	86137	24927	4→0	5118→5037	0	0	0	322079→0	337→0	0
Reegle	4	2	0	2013→1287	842→0	1	0	1718→0	317→0	2
SNOMED	0	1	0	119→115	0	0	0	60396→0	1234→0	0
SSW	6	1	0	118→46	22→0	0	0	723→0	4→0	0
STW	70	141	0	5004→5000	0	2	0	18533→0	5→0	0
UMBEL	2936	86	5→0	0	36535→0	0	0	740→0	0	0
UNESCO	0	1	0	19	0	0	0	124→0	0	0

Table 3.8. Validation results using the qSKOS quality analysis toolkit, part 3: *Linked Data Specific Issues*. Values marked with an asterisk (*) have been extrapolated from a randomly sampled subset of the concepts.

	Missing In-links	Missing Out-links	Broken Links	Undefined SKOS Resources	HTTP URI Scheme Violation
AGROVOC	31680*	17286	160*	0	0
DBpedia	865566*	865902	11400*	0	0
DDC	250790*	458	110*	0	0
EARTh	14349	9558	410	0	0
Eurovoc	6170*	6797	120790*	0	0
GEMET	3290*	584	40*	0	0
GeoNames	24	680	11	0	0
GTAA	171990*	171991	740*	0	0
IPSV	4731	4732	1	1	0
IPTC	2061	933→2061	2	1	0
LCSH	408920*	347560	2640*	0	0
LVAk		13411		0	0
NYTL	1892*	0	1376*	0	0
NYTP	4965	0	9	0	0
ODT	111	31	37	1	0
Plant	3246	0	662	0	0
PXV	1686	1046	107	0	0
RAMEAU	207260*	34803	132333*	0	0
Reegle	1447	809	321	1	9
SNOMED	102610*	0	5*	0	0
SSW	1941	1606	285	1	1→4
STW	6781	1463	504	0	0
UMBEL	26110*	0	130*	0	0
UNESCO	2509	2509	1	0	0

3.3.4 Recommendations for Best Practices

Many of the identified quality issues in SKOS vocabularies could have been prevented if the vocabulary publishers had been given clear guidelines on how to create and publish a good SKOS vocabulary. In particular, the question of what relationships to explicitly assert in the published vocabulary and what to leave for the vocabulary user to infer is not always clear. In practice, inference is not always possible or desirable for vocabulary users. Applications making use of SKOS vocabularies may benefit from explicitly asserted relations, even if they are in principle redundant and could have been inferred. In Publication V, the following guidelines for the inclusion of SKOS relationships in vocabularies published on the Web of Data are proposed:

1. Explicitly declare the types of SKOS `Concept`, `ConceptScheme` and `Collection` instances, even if they could be inferred. This is in line with the recommendation by Abdul Manaf et al. [3].

2. Include one or more concept schemes describing your vocabulary and label them appropriately. Assert the full set of both `topConceptOf` and `hasTopConcept` relationships. Make sure `inScheme` relationships are asserted for every concept.

3. Assert the full set of both `broader` and `narrower` relationships. This is also in line with the recommendation by Abdul Manaf et al. [3]. However, do not include the `broaderTransitive` and `narrowerTransitive` relationships, as they are only likely to be useful in special scenarios, may add a lot of new assertions to the vocabulary, and may be inferred by the vocabulary user when necessary.

4. Assert `related` properties both ways.

5. Assert mapping relationships only one way, with concepts from your own vocabulary as the subjects. This is to avoid "SKOS vocabulary hijacking", i.e., the assertion of facts about vocabularies published by others, which is similar to *ontology hijacking* [42].

4. Discussion and Conclusions

4.1 Research Questions Revisited

The research questions addressed in this dissertation were originally listed in Section 1:

1. How can user-friendly search facets be constructed for semantic portals which are based on complex hierarchical vocabularies?

2. What is the quality of automatically assigned subjects for documents written in inflected languages compared with subjects assigned by human indexers?

3. How can the technical quality and validity of controlled vocabularies expressed in SKOS format be automatically measured and improved?

To answer the first research question, we provided a method for creating user-centric facets for a semantic portal in publication I and evaluated its usability in publications I and II. The faceted search was implemented in an online prototype[1] of the HEALTHFINLAND system, which was awarded the 3rd prize[2] at the Semantic Web Challenge 2008. The work was subsequently incorporated into the production version of the HEALTHFINLAND portal[3], developed by the Finnish Institute for Health and Welfare.

To answer the second research question, we conducted experiments showing that the combination of a lemmatizer with the Maui toolkit provides

[1] http://demo.seco.tkk.fi/tervesuomi/
[2] http://challenge.semanticweb.org/submissions.html
[3] http://www.tervesuomi.fi

automatic subject indexing capabilities that are nearly as good as subjects assigned by human indexers in publication III. The work resulted in the ARPA information extraction toolkit, which was subsequently incorporated into the back-end system that is used to build CultureSampo [76], BookSampo [75], TravelSampo [78] and other recent semantic portal projects at the Semantic Computing Research Group.

Finally, to answer the first research question, we synthesized a list of quality criteria for SKOS vocabularies, analyzed in total 33 publicly available vocabularies, and attempted to correct as many problems as possible using the Skosify tool, in Publications IV and V. We found that nearly all vocabularies violated the SKOS integrity constraints, but we were able to automatically correct the great majority of such problems with our Skosify tool. Skosify has been released[4] as open source software under the MIT License. An online version of the tool is also available[5]. Publication IV was given the *Best In-use Paper*[6] award at the EKAW 2012 conference.

4.2 Research Evaluation

The research contained in this dissertation has been performed following the design science methodology in information systems research [37, 91], although this methodological background has not always been stated clearly in the publications.

To evaluate the research and show that it is consistent with the design science research methodology, we have analyzed the research according to Peffers et al's process model [91], in particular how well 1) the design problem is approached, 2) the problem is identified and motivated, 3) the objectives for the solution are defined, 4) the necessary design has been performed and the solution developed, 5) the designed solution has been demonstrated, 6) the solution has been evaluated, 7) the results have been communicated to both academic and non-technical audiences, and finally 8) what the core contributions of the research are. This analysis is summarized in Table 4.1.

[4] `http://code.google.com/p/skosify/`
[5] `http://demo.seco.tkk.fi/skosify`
[6] `http://ekaw2012.ekaw.org/awards`

Table 4.1. Analysis of research presented in this dissertation according to Peffers et al's [91] design science research methodology.

	User-centric Facets	**Automatic Subject Indexing**	**Vocabulary Quality**
Problem Identification and Motivation	Closing the gap between expert-oriented indexing vocabularies and needs of layman users [49], **I**, **II**	Automatic subject indexing of documents in highly inflected language for use in creating metadata for semantic portals **III**	Quality issues in controlled vocabularies that restrict their usability in semantic portals and other systems **IV**,**V**
Objectives of the Solution	Intuitive faceted search user interface which can be used to perform searches over existing health information web documents with metadata that references controlled vocabularies	As good automatic subject annotation of Finnish documents as has been attained with English, French and Spanish documents	Automatically identify and correct as many problems in SKOS vocabularies as possible
Design and Development	Facet categorization **I** Paper prototypes **I** Prototype system **I**, **II** Production system **II**	Combined state-of-the-art Maui tool [71] with three different stemmers and lemmatizers **III**	Quality criteria **IV**,**V** Skosify tool **IV**,**V**
Demonstration	Prototype system **I**, **II** Production system **II**	Tests on datasets from 2 different domains **III**	Tests on 33 vocabularies **IV**,**V** Online version **IV**
Evaluation	Closed card sort **I** Expert review of draft categorization **I** User tests on paper prototypes **I** User tests on prototype system **I**, **II**	Comparison of stemming and lemmatization strategies **III** Comparison of human vs. machine subject indexing **III** Test on datasets of another domain **III**	Evaluation of vocabularies before and after processing **IV**,**V** Performance evaluation **IV** Continuing use in ONKI system **IV**,**V**
Communication	Academic publications **I**, **II**, [49, 44] Non-technical publications Presentations by author and collaborators Online prototype and production systems Semantic Web Challenge 2008 3rd prize	Academic publication **III** Presentations by author and collaborators	Academic publications **IV**,**V** Code released as open source Mailing list announcements Presentations by author and collaborators Best In-use Paper award at EKAW 2012
Contribution	Method for creating user-friendly facets Demonstration of user-centric faceted search over existing web documents labelled with metadata that references controlled vocabularies Health information portal for the general public	Demonstration that using a lemmatizer improves automatic subject indexing Method incorporated into ARPA toolkit, used in several semantic applications	Quality criteria Finding that most vocabularies contain structural errors Method and tool for automatically correcting problems in vocabularies Best practice recommendations for vocabulary publishers Suggested amendment to SKOS specification

4.3 Limitations and Future Work

The method for creating user-centric facets has, to our knowledge, so far only been applied within the HEALTHFINLAND portal project. It has, however, been referred to as an example of user-centered facet design in other publications (e.g., [23, 24, 34, 39]). Whether the method is applicable or useful in other contexts is therefore not known. The card sorting and usability evaluations were performed with a relatively small number of participants (in total 12 participants in the card sorting, 8 participants in the usability evaluations). The number of participants in consistent with the *discount usability engineering* philosophy [56, 82] and roughly in line with recommendations on the number of participants in a card sorting experiment (Maurer and Warfel recommend 7–10 participants [70] while Nielsen recommends 15 [83]). However, such a small number of participants does not allow robust quantitative measures to be used for evaluating the outcome of the experiments.

The problems with unintuitive lower level categories that was caused by their reliance on the underlying vocabularies could possibly be avoided by using user-centric methods to design the lower category levels as well. This way, the categorization would become more like the ones used in faceted search systems where facets have been designed specifically for the system, such as SWED [99]. Naturally, this would take more work than simply reusing structures from the original vocabulary, and also the mappings to the underlying vocabulary would become more complex. Testing this approach was left for future work.

The automatic subject indexing experiments were all performed on documents and vocabularies in the Finnish language, with the assumption that the results would generalize to other inflected languages as well, i.e. that using a lemmatizer instead of a stemmer would improve the results of automatic subject indexing in languages such as Estonian, Turkish, Arabic, and Slavic languages. However, testing the approach on text in inflected languages other than Finnish is left for future work.

The vocabulary quality experiments were performed on a selection of 33 vocabularies. However, there are at least several hundred SKOS vocabularies available on the Web [4] likely having different quality attributes and varying levels of validity. An even wider and more systematic selection of vocabularies could reveal further problems in either the vocabularies themselves or in the vocabulary evaluation tools.

Furthermore, the focus of the quality evaluation was on computable, data-oriented quality issues, leaving out more intellectual quality criteria such as the applicability of a vocabulary for a particular purpose. Some of the correction heuristics, such as the removal of cycles, may cause an issue to be technically resolved, but the correction can be rather arbitrary and may not be the best possible action to take. A future study could compare the algorithmic corrections to corrections involving human judgment.

Bibliography

[1] ISO 2788: Guidelines for the establishment and development of monolingual thesauri. International Organization for Standardization (ISO), 1986.

[2] Structured vocabularies for information retrieval. Part 2: Thesauri (BS 8723-2:2005). British Standards Institution, 2005.

[3] ABDUL MANAF, N. A., BECHHOFER, S., AND STEVENS, R. Common modelling slips in SKOS vocabularies. In *Proceedings of the W3C Web Ontology Language (OWL) - Experiences and Directions Workshop (OWLED)* (2012), vol. 849 of *CEUR Workshop Proceedings*, CEUR-WS.org.

[4] ABDUL MANAF, N. A., BECHHOFER, S., AND STEVENS, R. The current state of SKOS vocabularies on the Web. In *Proceedings of the 9th Extended Semantic Web Conference (ESWC2012)*, vol. 7295 of *Lecture Notes in Computer Science*. Springer, 2012, pp. 270–284.

[5] AITCHISON, J., GILCHRIST, A., AND BAWDEN, D. *Thesaurus construction and use: a practical manual*. Aslib IMI, 2000.

[6] ALKULA, R. From plain character strings to meaningful words: Producing better full text databases for inflectional and compounding languages with morphological analysis software. *Information Retrieval 4*, 3 (2001), 195–208.

[7] BATTLE, L. Preliminary inventory of users and tasks for the semantic web. In *Proceedings of the 3rd International Semantic Web User Interaction Workshop (SWUI2006), Athens, GA, USA* (2006).

[8] BECHHOFER, S., AND MILES, A. SKOS Simple Knowledge Organization System reference. W3C recommendation, W3C, Aug. 2009. `http://www.w3.org/TR/2009/REC-skos-reference-20090818/`.

[9] BERNARD, M. Examining the effects of hypertext shape on user performance. *Usability News 4*, 2 (2002).

[10] BERNDTSSON, J. Designing an intranet from scratch to sketch: Experiences from techniques used in the IDEnet project. In *Proceedings of the 32nd Annual Hawaii International Conference on System Sciences - Volume 2* (1999), IEEE Computer Society, pp. 2019–.

[11] BERNERS-LEE, T., HENDLER, J., AND LASSILA, O. The Semantic Web. *Scientific American 284*, 5 (May 2001), 34–43.

[12] Berrueta, D., Fernández, S., and Frade, I. Cooking HTTP content negotiation with Vapour. In *Proceedings of the 4th Workshop on Scripting for the Semantic Web (SFSW2008)* (2008), vol. 368 of *CEUR Workshop Proceedings*, CEUR-WS.org.

[13] Bleier, A., Zapilko, B., Thamm, M., and Mutschke, P. Using SKOS to integrate social networking sites with scholarly information portals. In *SDoW2011 Social Data on the Web* (2011), vol. 830 of *CEUR Workshop Proceedings*, CEUR-WS.org.

[14] Borst, T., Fingerle, B., Neubert, J., and Seiler, A. How do libraries find their way onto the Semantic Web? *Liber Quarterly 19*, 3/4 (2010), 336–343.

[15] Buitelaar, P., and Declerck, T. Linguistic Annotation for the Semantic Web. In *Annotation for the Semantic Web*. IOS Press, Amsterdam, the Netherlands, 2003, pp. 93–110.

[16] Byrne, G., and Goddard, L. The strongest link: Libraries and linked data. *D-Lib Magazine 16*, 11/12 (2010).

[17] Carroll, J. J., and Klyne, G. Resource Description Framework (RDF): Concepts and abstract syntax. W3C recommendation, W3C, Feb. 2004. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/.

[18] Cowie, J., and Lehnert, W. Information extraction. *Communications of the ACM 39*, 1 (Jan. 1996), 80–91.

[19] Şah, M., Hall, W., Gibbins, N. M., and De Roure, D. C. Semport: a personalized semantic portal. In *Proceedings of the 18th conference on Hypertext and Hypermedia (HT'07)* (2007), ACM, pp. 31–32.

[20] Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. GATE: an architecture for development of robust HLT applications. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02)* (2002), Association for Computational Linguistics, pp. 168–175.

[21] Cutter, C., Cutter, W., Ford, W., Phillips, P., and Conneck, O. *Rules for a Dictionary Catalog*. Special report on public libraries. U.S. Government Printing Office, 1904.

[22] Dewey, M. *A Classification and Subject Index For Cataloguing and Arranging the Books and Pamphlets of a Library*. Cass, Lockwood & Brainard Company, 1876.

[23] Ding, Y., Sun, Y., Chen, B., Borner, K., Ding, L., Wild, D., Wu, M., DiFranzo, D., Fuenzalida, A., Li, D., Milojevic, S., Chen, S., Sankaranarayanan, M., and Toma, I. Semantic web portal: A platform for better browsing and visualizing semantic data. In *Active Media Technology*, vol. 6335 of *Lecture Notes in Computer Science*. Springer, 2010, pp. 448–460.

[24] Dong, H., and Hussain, F. K. Semantic service matchmaking for digital health ecosystems. *Knowledge-Based Systems 24*, 6 (2011), 761 – 774.

[25] El-Shishtawy, T., and Al-Sammak, A. Arabic keyphrase extraction using linguistic knowledge and machine learning techniques. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools* (Cairo, Egypt, April 2009), The MEDAR Consortium.

[26] Elliott, A. M. *Computational support for sketching and image sorting during the early phase of architectural design*. PhD thesis, University of California at Berkeley, 2002.

[27] Fürber, C., and Hepp, M. Using Semantic Web resources for data quality management. In *Proceedings of the 17th International Conference on Knowledge Engineering and Knowledge Management (EKAW2010)*, vol. 6317 of *Lecture Notes in Computer Science*. Springer, 2010, pp. 211–225.

[28] Giunchiglia, F., Marchese, M., and Zaihrayeu, I. Encoding classifications into lightweight ontologies. In *Proceedings of the 3rd European Semantic Web Conference (ESWC2006)*, vol. 4011 of *Lecture Notes in Computer Science*. Springer, 2006, pp. 80–94.

[29] Greenberg, J., Losee, R., Agüera, J. R. P., Scherle, R., White, H., and Willis, C. HIVE: Helping interdisciplinary vocabulary engineering. *Bulletin of the American Society for Information Science and Technology 37*, 4 (2011), 23–26.

[30] Gruber, T. R. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies 43*, 5-6 (Dec. 1995), 907–928.

[31] Grüninger, M., and Fox, M. S. Methodology for the design and evaluation of ontologies. In *Proceedings of the IJCAI95 Workshop on Basic Ontological Issues in Knowledge Sharing* (1995).

[32] Guarino, N., and Welty, C. Evaluating ontological decisions with OntoClean. *Communications of the ACM 45*, 2 (Feb. 2002), 61–65.

[33] Guha, R. V., and Brickley, D. RDF vocabulary description language 1.0: RDF Schema. W3C recommendation, W3C, Feb. 2004. `http://www.w3.org/TR/2004/REC-rdf-schema-20040210/`.

[34] Gustafsson, M. *SOMWeb: supporting a distributed clinical community of practice using semantic web technologies*. PhD thesis, Chalmers University of Technology, 2009.

[35] Hearst, M., Elliott, A., English, J., Sinha, R., Swearingen, K., and Lee, K.-P. Finding the flow in web site search. *Communications of the ACM 45*, 9 (2002), 42–49.

[36] Heath, T., and Bizer, C. *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, 2011.

[37] Hevner, A. R., March, S. T., Park, J., and Ram, S. Design science in information systems research. *MIS Quarterly 28*, 1 (Mar. 2004), 75–105.

[38] Hildebrand, M., Ossenbruggen, J., and Hardman, L. An analysis of search-based user interaction on the semantic web. Tech. Rep. INS-E0706, Centrum voor Wiskunde en Informatica (CWI), 2007. `http://oai.cwi.nl/oai/asset/12302/12302D.pdf`.

[39] Hildebrand, M., and Van Ossenbruggen, J. Configuring Semantic Web interfaces by data mapping. In *Proceedings of the IUI'09 Workshop on Visual Interfaces to the Social and the Semantic Web (VISSW 2009)* (Februrary 2009), vol. 443 of *CEUR Workshop Proceedings*, CEUR-WS.org.

[40] Hildebrand, M., van Ossenbruggen, J., and Hardman, L. /facet: A browser for heterogeneous Semantic Web repositories. In *Proceedings of the 5th International Semantic Web Conference (ISWC2006)*, vol. 4273 of *Lecture Notes in Computer Science*. Springer, 2006, pp. 272–285.

[41] Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S., and Pylkkönen, J. Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech & Language 20*, 4 (2006), 515–541.

[42] Hogan, A., Harth, A., Passant, A., Decker, S., and Polleres, A. Weaving the pedantic web. In *Proceedings of the WWW2010 Workshop on Linked Data on the Web (LDOW2010)* (2010), vol. 628 of *CEUR Workshop Proceedings*, CEUR-WS.org.

[43] Horridge, M., Parsia, B., and Sattler, U. Explaining inconsistencies in OWL ontologies. In *Proceedings of the 3rd International Conference on Scalable Uncertainty Management (SUM '09)* (2009), vol. 5785 of *Lecture Notes in Computer Science*, Springer, pp. 124–137.

[44] Hyvönen, E., Viljanen, K., Suominen, O., and Hukka, E. HealthFinland – publishing health promotion information on the Semantic Web. In *Proceedings of DrMED 2008: International Workshop on Describing Medical Web Resources. The 21st International Congress on the European Federation for Medical Informatics (MIE 2008), Göteborg, Sweden* (May 2008).

[45] Hyvönen, E., Viljanen, K., Tuominen, J., and Seppälä, K. Building a national Semantic Web ontology and ontology service infrastructure—the FinnONTO approach. In *Proceedings of the 5th European Semantic Web Conference (ESWC2008)*, vol. 5021 of *Lecture Notes in Computer Science*. Springer, 2008, pp. 95–109.

[46] Hyvönen, E., Mäkelä, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., Junnila, M., and Kettula, S. MuseumFinland – Finnish museums on the Semantic Web. *Journal of Web Semantics 3*, 2 (2005), 25.

[47] Hyvönen, E., Styrman, A., and Saarela, S. Ontology-based image retrieval. In *Towards the Semantic Web and Web Services, Proceedings of XML Finland 2002 Conference* (Helsinki, Finland, October 21–22 2002), HIIT Publications, pp. 15–27.

[48] Hyvönen, E., Viljanen, K., Mäkelä, E., Kauppinen, T., Ruotsalo, T., Valkeapää, O., Seppälä, K., Suominen, O., Alm, O., Lindroos, R., Känsälä, T., Henriksson, R., Frosterus, M., Tuominen, J., Sinkkilä, R., and Kurki, J. Elements of a national Semantic Web infrastructure - case study Finland on the Semantic Web. In *Proceedings of the First International Semantic Computing Conference (IEEE ICSC 2007), Irvine, California* (September 2007), IEEE Press, pp. 216–223.

[49] Hyvönen, E., Viljanen, K., and Suominen, O. HealthFinland–Finnish health information on the Semantic Web. In *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC2007 + ASWC2007)*, vol. 4825 of *Lecture Notes in Computer Science*. Springer, 2007, pp. 778–791.

[50] Kalyanpur, A. *Debugging and repair of OWL ontologies*. PhD thesis, University of Maryland at College Park, 2006.

[51] Kettunen, K., Kunttu, T., and Järvelin, K. To stem or lemmatize a highly inflectional language in a probabilistic IR environment? *Journal of Documentation 61*, 4 (2005), 476–496.

[52] Kiryakov, A., Popov, B., Terziev, I., Manov, D., and Ognyanoff, D. Semantic annotation, indexing, and retrieval. *Web Semantics: Science, Services and Agents on the World Wide Web 2*, 1 (2004), 49–79.

[53] Kless, D., and Milton, S. Towards quality measures for evaluating thesauri. In *Metadata and Semantic Research*, vol. 108 of *Communications in Computer and Information Science*. Springer, 2010, pp. 312–319.

[54] Korenius, T., Laurikkala, J., Järvelin, K., and Juhola, M. Stemming and lemmatization in the clustering of Finnish text documents. In *Proceedings of the 13th ACM international conference on Information and knowledge management (CIKM'04)* (New York, NY, USA, 2004), ACM, pp. 625–633.

[55] Koskenniemi, K. *Two-level Morphology: A General Computational Model*. PhD thesis, University of Helsinki, 1983.

[56] Krug, S. *Don't Make Me Think: A Common Sense Approach to Web Usability*, second ed. New Riders Press, August 2005.

[57] Känsälä, T., and Hyvönen, E. A semantic view-based portal utilizing Learning Object Metadata. In *Proceedings of the 1st Asian Semantic Web Conference (ASWC2006), Semantic Web Applications and Tools Workshop.* (Beijing, China, August 2006).

[58] Lamantia, J. Analyzing card sort results with a spreadsheet template. *Boxes and Arrows* (Aug 26 2003). `http://boxesandarrows.com/S1708`.

[59] Lauesen, S. *User Interface Design: A Software Engineering Perspective*. Pearson/Addison-Wesley, Harlow, England, 2005.

[60] Lausen, H., Ding, Y., Stollberg, M., Fensel, D., Hernandez, R., and Han, S. Semantic web portals: state-of-the-art survey. *Journal of Knowledge Management 9*, 5 (2005), 40–49.

[61] Leventhal, L., and Barnes, J. *Usability engineering: Process, products and examples*. Prentice Hall, 2007.

[62] Lindén, K., Silfverberg, M., and Pirinen, T. HFST tools for morphology – an efficient open-source package for construction of morphological analyzers. In *State of the Art in Computational Morphology*, vol. 41 of *Communications in Computer and Information Science*. Springer, 2009, pp. 28–47.

[63] Lovins, J. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics 11* (1968), 22–31.

[64] Löfberg, L., Piao, S., Nykanen, A., Varantola, K., Rayson, P., and Juntunen, J.-P. A semantic tagger for the Finnish language. In *Proceedings of the Corpus Linguistics 2005 Conference* (Birmingham, UK, July 2005), University of Birmingham, Centre for Corpus Research.

[65] Mader, C., Haslhofer, B., and Isaac, A. Finding quality issues in SKOS vocabularies. In *Theory and Practice of Digital Libraries*, vol. 7489 of *Lecture Notes in Computer Science*. Springer, 2012, pp. 222–233.

[66] Maedche, A., Staab, S., Stojanovic, N., Studer, R., and Sure, Y. SEmantic portAL - the SEAL approach. In *Spinning the Semantic Web*. MIT Press, 2001, pp. 317–359.

[67] Malmsten, M. Making a library catalogue part of the semantic web. In *Proceedings of the International Conference on Dublin Core and Metadata Applications (DC-2008)* (Berlin, Germany, Sept. 2008), Dublin Core Metadata Initiative, pp. 146–152.

[68] Marchionini, G. Exploratory search: from finding to understanding. *Communications of the ACM 49*, 4 (2006), 41–46.

[69] Maron, M. E. Automatic indexing: an experimental inquiry. *Journal of the ACM (JACM) 8*, 3 (1961), 404–417.

[70] Maurer, D., and Warfel, T. Card sorting: a definitive guide. *Boxes and Arrows* (Apr 7 2003). http://boxesandarrows.com/S1937.

[71] Medelyan, O. *Human-competitive automatic topic indexing*. PhD thesis, University of Waikato, Department of Computer Science, 2009.

[72] Medelyan, O., and Witten, I. H. Thesaurus based automatic keyphrase indexing. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries (JCDL'06)* (New York, NY, USA, 2006), ACM, pp. 296–297.

[73] Miles, A., Rogers, N., and Beckett, D. Migrating thesauri to the Semantic Web – Guidelines and case studies for generating RDF encodings of existing thesauri. SWAD-Europe project deliverable 8.8, SWAD-Europe, 2004. http://www.w3.org/2001/sw/Europe/reports/thes/8.8/.

[74] Morville, P. *Ambient Findability: What We Find Changes Who We Become*. O'Reilly Media, Inc., 2005.

[75] Mäkelä, E., Hypén, K., and Hyvönen, E. Improving fiction literature access by Linked Open Data -based collaborative knowledge storage - the BookSampo project. In *Proceedings of the World Library and Information Congress: 78th IFLA General Conference and Assembly* (Helsinki, Finland, Aug. 2012), IFLA.

[76] Mäkelä, E., Hyvönen, E., and Ruotsalo, T. How to deal with massively heterogeneous cultural heritage data – lessons learned in CultureSampo. *Semantic Web – Interoperability, Usability, Applicability 3*, 1 (2012).

[77] Mäkelä, E., Hyvönen, E., and Sidoroff, T. View-based user interfaces for information retrieval on the Semantic Web. In *Proceedings of the ISWC2005 Workshop on End User Semantic Web Interaction (SWUI2005)* (Nov 2005).

[78] Mäkelä, E., Lindblad, A., Väätäinen, J., Alatalo, R., Suominen, O., and Hyvönen, E. Discovering places of interest through direct and indirect associations in heterogeneous sources — the TravelSampo system. In *Terra Cognita 2011: Foundations, Technologies and Applications of the Geospatial Web* (2011), vol. 798 of *CEUR Workshop Proceedings*, CEUR-WS.org.

[79] Nadeau, D., and Sekine, S. A survey of named entity recognition and classification. *Lingvisticae Investigationes 30*, 1 (2007), 3–26.

[80] Nagy, H., Pellegrini, T., and Mader, C. Exploring structural differences in thesauri for SKOS-based applications. In *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics '11)* (Graz, Austria, 2011), ACM, pp. 187–190.

[81] Neubert, J. Bringing the "Thesaurus for Economics" on to the Web of Linked Data. In *Proceedings of the WWW2009 Workshop on Linked Data on the Web (LDOW2009)* (2009), vol. 538 of *CEUR Workshop Proceedings*, CEUR-WS.org.

[82] Nielsen, J. Guerrilla HCI: using discount usability engineering to penetrate the intimidation barrier. In *Cost-justifying usability*. Academic Press, Inc., Orlando, FL, USA, 1994, pp. 245–272.

[83] Nielsen, J. Card sorting: How many users to test, Jul 19 2004. Alertbox column, `http://www.useit.com/alertbox/20040719.html`.

[84] Nielsen, J., and Sano, D. SunWeb: User interface design for Sun Microsystem's internal web. In *Proceedings of the 2nd World Wide Web Conference, Chicago, IL* (Oct 17-20 1994), pp. 547–557.

[85] Oflazer, K., and Kuruöz, I. Tagging and morphological disambiguation of Turkish text. In *Proceedings of the fourth conference on Applied natural language processing (ANLC '94)* (Stuttgart, Germany, 1994), Association for Computational Linguistics, pp. 144–149.

[86] Oren, E., Delbru, R., and Decker, S. Extending faceted navigation for RDF data. In *Proceedings of the 5th International Semantic Web Conference (ISWC2006)*, vol. 4273 of *Lecture Notes in Computer Science*. Springer, 2006, pp. 559–572.

[87] Otlet, P. *Traité de documentation. Le livre sur le livre. Théorie et pratique*. No. 197 in IIB Publication. Editiones Mundaneum, Brussels, 1934.

[88] Ovchinnikova, E., Wandmacher, T., and Kühnberger, K. Solving terminological inconsistency problems in ontology design. *International Journal of Interoperability in Business Information Systems 2*, 1 (2007), 65–80.

[89] Pala, N., and Çiçekli, I. Turkish keyphrase extraction using KEA. In *Proceedings of the 22nd International Symposium on Computer and Information Sciences (ISCIS 2007)* (Ankara, Turkey, Nov. 2007), pp. 1–5.

[90] Passin, T. B. *Explorer's Guide to the Semantic Web*. Manning, 2004.

[91] Peffers, K., Tuunanen, T., Rothenberger, M., and Chatterjee, S. A design science research methodology for information systems research. *Journal of Management Information Systems 24*, 3 (2007), 45–77.

[92] Pennanen, P., and Alatalo, T. Leiki – a platform for personalized content targeting. In *Proceedings of the 12th ACM conference on Hypertext and Hypermedia (HYPERTEXT'01)* (2001), ACM, p. 6.

[93] Pollitt, A., Ellis, G., and Smith, M. HIBROWSE for bibliographic databases. *Journal of information science 20*, 6 (1994), 413–426.

[94] Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D., and Kirilov, A. KIM–a semantic platform for information extraction and retrieval. *Natural language engineering 10*, 3-4 (2004), 375–392.

[95] Porter, M. An algorithm for suffix stripping. *Program 14* (1980), 130–137.

[96] Poveda-Villalón, M., Suárez-Figueroa, M., and Gómez-Pérez, A. Validating ontologies with OOPS! In *Proceedings of the 18th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2012)*, vol. 7603 of *Lecture Notes in Computer Science*. Springer, 2012, pp. 267–281.

[97] Ranganathan, S. *The colon classification*, vol. 4. Graduate School of Library Service, Rutgers, the State University, 1965.

[98] Rector, A., Bechhofer, S., Goble, C., Horrocks, I., Nowlan, W., and Solomon, W. The GRAIL concept modelling language for medical terminology. *Artificial Intelligence in Medicine 9*, 2 (1997), 139 – 171.

[99] Reynolds, D., Shabajee, P., and Cayzer, S. Semantic information portals. In *Proceedings of the 13th International World Wide Web Conference, Alternate track papers & posters (WWW Alt. '04)* (New York, May 2004), ACM, pp. 290–291.

[100] Rosenfeld, L., and Morville, P. *Information Architecture for the World Wide Web*, third ed. O'Reilly, 2006.

[101] Rugg, G., and McGeorge, P. The sorting techniques: a tutorial paper on card sorts, picture sorts and item sorts. *Expert Systems 14*, 2 (1997), 80–93.

[102] Salton, G., and Buckley, C. Term-weighting approaches in automatic text retrieval. *Information Processing and Management 24*, 5 (1988), 513–523.

[103] Schandl, T., and Blumauer, A. PoolParty: SKOS thesaurus management utilizing Linked Data. In *Proceedings of the 7th Extended Semantic Web Conference (ESWC2010)*, vol. 6089 of *Lecture Notes in Computer Science*. Springer, 2010, pp. 421–425.

[104] schraefel, m., Karam, M., and Zhao, S. mSpace: interaction design for user-determined, adaptable domain exploration in hypermedia. In *Proceedings of AH 2003: Workshop on Adaptive Hypermedia and Adaptive Web Based Systems* (Nottingham, UK, 2003).

[105] SCHRAEFEL, M., SMITH, D. A., RUSSEL, A., OWENS, A., HARRIS, C., AND WILSON, M. The mSpace Classical Music Explorer: Improving access to classical music for real people. In *Proceedings of V MUSICNETWORK OPEN WORKSHOP: Integration of Music in Multimedia Applications* (Vienna, Austria, 2005).

[106] SCHREIBER, G., AMIN, A., ASSEM, M., BOER, V., HARDMAN, L., HILDEBRAND, M., HOLLINK, L., HUANG, Z., KERSEN, J., NIET, M., OMELAYENKO, B., OSSENBRUGGEN, J., SIEBES, R., TAEKEMA, J., WIELEMAKER, J., AND WIELINGA, B. MultimediaN E-Culture demonstrator. In *Proceedings of the 5th International Semantic Web Conference (ISWC2006)*, vol. 4273 of *Lecture Notes in Computer Science*. Springer, 2006, pp. 951–958.

[107] SCHREIBER, G., AND DEAN, M. OWL Web Ontology Language reference. W3C recommendation, W3C, Feb. 2004. http://www.w3.org/TR/2004/REC-owl-ref-20040210/.

[108] SEBASTIANI, F. Machine learning in automated text categorization. *ACM Computing Surveys 34*, 1 (2002), 1–47.

[109] SHADBOLT, N., AND BURTON, M. Knowledge elicitation. In *Evaluation of Human Work: A Practical Ergonomics Methodology*. Taylor and Francis, 1990, pp. 321–346.

[110] SHADBOLT, N., GIBBINS, N., GLASER, H., HARRIS, S., AND M.C. SCHRAEFEL. CS AKTive Space, or how we learned to stop worrying and love the Semantic Web. *IEEE Intelligent Systems 19*, 3 (2004), 41–47.

[111] SIDOROFF, T., AND HYVÖNEN, E. Semantic e-goverment portals - a case study. In *Proceedings of the ISWC-2005 Workshop Semantic Web Case Studies and Best Practices for eBusiness SWCASE05* (Nov 2005).

[112] SPENCER, D. *Card sorting: Designing usable categories*. Rosenfeld Media, 2009.

[113] STAAB, S., ANGELE, J., DECKER, S., ERDMANN, M., HOTHO, A., MAEDCHE, A., SCHNURR, H.-P., STUDER, R., AND SURE, Y. Semantic community web portals. *Computer Networks 33*, 1–6 (2000), 473 – 491.

[114] STEVENS, R., GOBLE, C. A., AND BECHHOFER, S. Ontology-based knowledge representation for bioinformatics. *Briefings in bioinformatics 1*, 4 (2000), 398–414.

[115] STOICA, E., AND HEARST, M. A. Nearly-automated metadata hierarchy creation. In *Proceedings of HLT-NAACL 2004: Short Papers* (Boston, Massachusetts, 2004), Association for Computational Linguistics, pp. 117–120.

[116] SUMMERS, E., ISAAC, A., REDDING, C., AND KRECH, D. LCSH, SKOS and Linked Data. In *Proceedings of the International Conference on Dublin Core and Metadata Applications (DC-2008)* (Berlin, Germany, Sept. 2008), Dublin Core Metadata Initiative, pp. 25–33.

[117] SUOMINEN, O. Käyttäjäkeskeinen moninäkymähaku semanttisessa portaalissa (user-centric faceted search in a semantic portal). Master's thesis, University of Helsinki, Dept. of Computer Science, February 2008.

[118] SUOMINEN, O., AND HYVÖNEN, E. Expressing and aggregating rich event descriptions. In *Proceedings of the 6th Workshop on Scripting and Development on the Semantic Web (SFSW 2010)* (Heraklion, Greece, May 2010), vol. 699 of *CEUR Workshop Proceedings*, CEUR-WS.org.

[119] SVENONIUS, E. *The Intellectual Foundation of Information Organization*. MIT Press, 2000.

[120] TAPANAINEN, P., AND JÄRVINEN, T. A non-projective dependency parser. In *Proceedings of the Fifth Conference on Applied Natural Language Processing* (1997).

[121] TORDAI, A., AND RIJKE, M. Four stemmers and a funeral: Stemming in Hungarian at CLEF 2005. In *Accessing Multilingual Information Repositories*, vol. 4022 of *Lecture Notes in Computer Science*. Springer, 2006, pp. 179–186.

[122] TRIESCHNIGG, D., PEZIK, P., LEE, V., DE JONG, F., KRAAIJ, W., AND REBHOLZ-SCHUHMANN, D. MeSH Up: Effective MeSH text classification for improved document retrieval. *Bioinformatics 25*, 11 (2009), 1412–1418.

[123] TUDHOPE, D., BINDING, C., BLOCKS, D., AND CUNLIFFE, D. Query expansion via conceptual distance in thesaurus indexed collections. *Journal of Documentation 62*, 4 (2006), 509–533.

[124] TULLIS, T. Intranet organization is in the cards. *Journal of Intranet Strategy and Management 1*, 1 (2003), 5–8.

[125] TUNKELANG, D. Faceted search. *Synthesis Lectures on Information Concepts, Retrieval, and Services 1*, 1 (2009), 1–80.

[126] TUOMINEN, J., FROSTERUS, M., VILJANEN, K., AND HYVÖNEN, E. ONKI SKOS server for publishing and utilizing SKOS vocabularies and ontologies as services. In *Proceedings of the 6th European Semantic Web Conference (ESWC2009)*, vol. 5554 of *Lecture Notes in Computer Science*. Springer, 2009, pp. 768–780.

[127] VALKEAPÄÄ, O., ALM, O., AND HYVÖNEN, E. Efficient content creation on the Semantic Web using metadata schemas with domain ontology services. In *Proceedings of the 4th European Semantic Web Conference (ESWC2007)*, vol. 4519 of *Lecture Notes in Computer Science*. Springer, 2007, pp. 819–828.

[128] VAN ASSEM, M., MALAISÉ, V., MILES, A., AND SCHREIBER, G. A method to convert thesauri to SKOS. In *Proceedings of the 3rd European Semantic Web Conference (ESWC2006)*, vol. 4011 of *Lecture Notes in Computer Science*. Springer, 2006, pp. 95–109.

[129] VEHVILÄINEN, A., HYVÖNEN, E., AND ALM, O. A semi-automatic semantic annotation and authoring tool for a library help desk service. In *Emerging Technologies for Semantic Work Environments: Techniques, Methods, and Applications*. IGI Group, Hershey, USA, 2008, pp. 100–114.

[130] VICKERY, B. *Faceted classification: a guide to construction and use of special schemes*, vol. 3. Aslib, London, 1960.

[131] VILJANEN, K., KÄNSÄLÄ, T., HYVÖNEN, E., AND MÄKELÄ, E. Ontodella - a projection and linking service for Semantic Web applications. In *Proceedings of the 17th International Conference on Database and Expert Systems Applications (DEXA 2006)* (Sept. 2006), IEEE, pp. 370–376.

[132] VOORHEES, E. M. Query expansion using lexical-semantic relations. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '94)* (Dublin, Ireland, 1994), Springer, pp. 61–69.

[133] VRANDECIC, D. *Ontology Evaluation*. PhD thesis, KIT, Fakultät für Wirtschaftswissenschaften, Karlsruhe, 2010.

[134] WANG, Y., SURE, Y., STEVENS, R., AND RECTOR, A. Knowledge elicitation plug-in for Protégé: Card sorting and laddering. In *Proceedings of the 1st Asian Semantic Web Conference (ASWC 2006)*, vol. 4185 of *Lecture Notes in Computer Science*. Springer, 2006, pp. 552–565.

[135] WITTEN, I. H., PAYNTER, G., FRANK, E., GUTWIN, C., AND NEVILL-MANNING, C. G. KEA: Practical automatic keyphrase extraction. In *Proceedings of Digital Libraries 99* (1999).

Faceted search user interfaces provide means to explore, e.g., items for sale, museum collections or recipes.A method for creating user-centric search facets that simplify complex hierarchical displays is presented.

Automatic subject indexing allows text documents to be tagged with semantic information, avoiding laborious manual work in libraries, museums and other organizations. A method is developed for automatically determining the themes of Finnish language text.

Controlled vocabularies such as taxonomies and thesauri form the heart of many semantic applications. A method for assessing and improving the *quality* of such artefacts is presented.

Cover image:
Great raft spider *(Dolomedes plantarius)*
Evitskog, Finland, 2 July 2013

ISBN 978-952-60-5253-3
ISBN 978-952-60-5254-0 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Media Technology
www.aalto.fi

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL DISSERTATIONS

Aalto-DD 113/2013