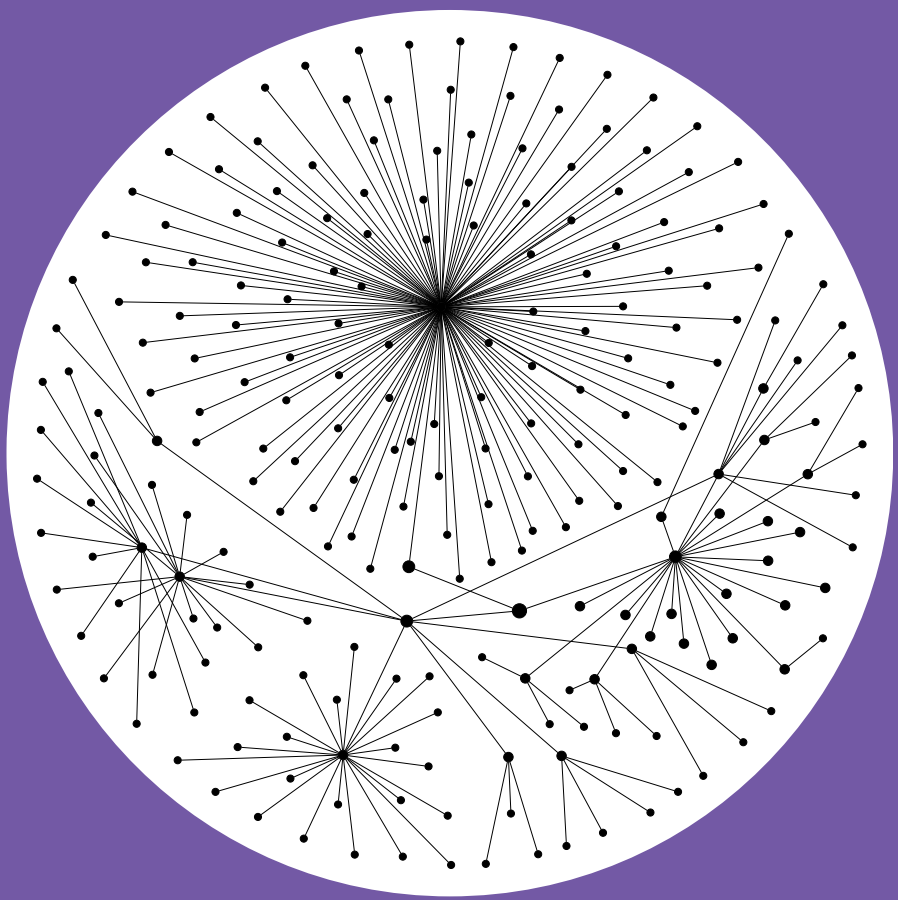
Department of Computer Science

Ontology Services for Knowledge Organization Systems

Jouni Tuominen



Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 101/2017

Ontology Services for Knowledge Organization Systems

Jouni Tuominen

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall TU1 of the school on 16 June 2017 at 12 noon.

Aalto University
School of Science
Department of Computer Science
Semantic Computing Research Group

Supervising professor
Eero Hyvönen

Thesis advisor
Eero Hyvönen

Preliminary examiners
Professor Vagan Terziyan, University of Jyväskylä, Finland
Professor Mathieu d'Aquin, National University of Ireland Galway, Ireland

Opponent
Professor Marcia Lei Zeng, Kent State University, United States

Aalto University publication series
DOCTORAL DISSERTATIONS 101/2017

ISBN 978-952-60-7456-6 (printed)
ISBN 978-952-60-7455-9 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-7455-9

Unigrafia Oy
Helsinki 2017

Finland

441 697
Printed matter

Aalto University

Abstract

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

Author
Jouni Tuominen

Name of the doctoral dissertation
Ontology Services for Knowledge Organization Systems

Publisher School of Science

Unit Department of Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 101/2017

Field of research Knowledge Technology

Manuscript submitted 12 March 2017 **Date of the defence** 16 June 2017

Permission to publish granted (date) 25 April 2017 **Language** English

☐ **Monograph** ☒ **Article dissertation** ☐ **Essay dissertation**

Abstract

Ontologies and other knowledge organization systems, such as controlled vocabularies, can be used to enhance the findability of information. By describing the contents of documents using a shared, harmonized terminology, information systems can provide efficient search and browsing functionalities for the contents. Explicit descriptive metadata aims to solve some of the prevailing issues in full text search in many search engines, including the processing of synonyms and homonyms. The use of ontologies as domain models enables the machine-processability of contents, semantic reasoning, information integration, and other intelligent ways of processing the data.

The utilization of knowledge organization systems in content indexing and information retrieval can be facilitated by providing automated tools for their efficient use. This thesis studies and presents novel methods and systems for publishing and using knowledge organization systems as ontology services. The research is conducted by designing and evaluating prototype systems that support the use of ontologies in real-life use cases. The research follows the principles of the design science and action research methodologies.

The presented ONKI system provides user interface components and application programming interfaces that can be integrated into external applications to enable ontology-based workflows. The features of the system are based on analyzing the needs of the main user groups of ontologies. The common functionalities identified in ontology-based workflows include concept search, browsing, and selection.

The thesis presents the Linked Open Ontology cloud approach for managing and publishing a set of interlinked ontologies in an ontology service. The system enables the users to use multiple ontologies as a single, interoperable, cross-domain representation instead of individual ontologies. For facilitating the simultaneous use of ontologies published in different ontology repositories, the Normalized Ontology Repository approach is presented.

As a use case of managing and publishing a semantically rich knowledge organization system as an ontology, the thesis presents the Taxon Meta-Ontology model for biological nomenclatures and classifications. The model supports the representation of changes and differing opinions of taxonomic concepts.

The ONKI system and the ontologies developed using the methods presented in this thesis have been provided as a living lab service http://onki.fi, which has been run since 2008. The service provides tools and support for the users of ontologies, including content indexers, information searchers, ontology developers, and application developers.

Keywords semantic web, knowledge organization systems, metadata creation, ontology services

ISBN (printed) 978-952-60-7456-6 **ISBN (pdf)** 978-952-60-7455-9

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

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2017

Pages 198 **urn** http://urn.fi/URN:ISBN:978-952-60-7455-9

Tekijä
Jouni Tuominen

Väitöskirjan nimi
Ontology Services for Knowledge Organization Systems

Julkaisija Perustieteiden korkeakoulu

Yksikkö Tietotekniikan laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 101/2017

Tutkimusala Tietämystekniikka

Käsikirjoituksen pvm 12.03.2017 **Väitöspäivä** 16.06.2017

Julkaisuluvan myöntämispäivä 25.04.2017 **Kieli** Englanti

☐ **Monografia** ☒ **Artikkeliväitöskirja** ☐ **Esseeväitöskirja**

Tiivistelmä

Ontologioita ja muita tietämyksen järjestämisen menetelmiä, kuten kontrolloituja sanastoja, voidaan käyttää tiedon löytämisen parantamiseksi. Kun dokumenttien sisällöt kuvaillaan käyttämällä jaettua, yhtenäistettyä terminologiaa, tietojärjestelmät voivat tarjota tehokkaita haku- ja selaustoiminnallisuuksia sisältöihin. Eksplisiittisesti esitetty, kuvaileva metatieto pyrkii ratkaisemaan monien hakukoneiden käyttämän kokotekstihaun ongelmia, kuten synonyymien ja homonyymien huomioimisen. Ontologioiden käyttäminen käsitemalleina mahdollistaa sisältöjen koneellisen käsittelyn, semanttisen päättelyn, tiedon integroinnin ja muita älykkäitä menetelmiä.

Tietämyksen järjestämisen menetelmien hyödyntämistä sisältöjen indeksoinnissa ja haussa voidaan helpottaa tarjoamalla käyttäjille automatisoituja työkaluja niiden tehokkaaseen käyttämiseen. Tässä väitöskirjassa tutkitaan ja esitetään uudenlaisia menetelmiä ja järjestelmiä tietämyksen järjestämisen menetelmien julkaisemiseksi ontologiapalveluina. Tutkimus on toteutettu suunnittelemalla ja arvioimalla prototyyppijärjestelmiä, jotka edistävät ontologioiden käyttämistä todellisissa käyttötapauksissa. Tutkimus nojautuu suunnittelutieteen ja toimintatutkimuksen metodologioiden periaatteisiin.

Työssä esitetty ONKI-järjestelmä tarjoaa käyttöliittymäkomponentteja ja ohjelmallisia rajapintoja, jotka voidaan integroida ulkoisiin sovelluksiin ontologiaperustaisten työnkulkujen mahdollistamiseksi. Järjestelmän ominaisuudet on toteutettu perustuen ontologioiden keskeisten käyttäjäryhmien tarpeiden selvittämiseen. Ontologiaperustaisista työnkuluista tunnistettuja yleisiä toiminnallisuuksia ovat käsitteen haku, selailu ja valinta.

Tässä työssä esitetään linkitetyn avoimen ontologiapilven menetelmä toisiinsa linkitettyjen ontologioiden ylläpitämiseen ja julkaisemiseen ontologiapalvelussa. Järjestelmän avulla käyttäjät voivat käyttää useita ontologioita yhtenä, yhteentoimivana, alat yhdistävänä kokonaisuutena erillisten ontologioiden sijaan. Eri ontologiapalveluissa julkaistujen ontologioiden samanaikaisen käytön helpottamiseksi esitetään normalisoidun ontologiapalvelun menetelmä.

Käyttötapauksena semanttisesti rikkaan tietämyksen järjestämisen menetelmän ylläpitämisestä ja julkaisemisesta työssä esitetään biologisten nimistöjen ja luokitusten taksonominen ontologiamalli. Malli mahdollistaa taksonomisten käsitteiden muutosten ja toisistaan poikkeavien näkemysten esittämisen.

ONKI-järjestelmä ja työssä esitetyillä menetelmillä kehitetyt ontologiat ovat olleet käytettävissä living lab -palvelussa http://onki.fi, joka on ollut toiminnassa vuodesta 2008 lähtien. Palvelu tarjoaa työkaluja ja tukea ontologioiden käyttäjille, kuten tiedon indeksoijille, hakijoille, ontologioiden kehittäjille ja sovelluskehittäjille.

Avainsanat semanttinen web, tietämyksen järjestämisen menetelmät, metadatan tuottaminen, ontologiapalvelut

ISBN (painettu) 978-952-60-7456-6 **ISBN (pdf)** 978-952-60-7455-9

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

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2017

Sivumäärä 198 **urn** http://urn.fi/URN:ISBN:978-952-60-7455-9

Preface

The seeds for this dissertation were planted in 2007 when I started working in the Semanting Computing Research Group (SeCo), at the Department of Computer Science, Aalto University (formerly the Department of Media Technology, Helsinki University of Technology) and the Department of Computer Science, University of Helsinki, Finland.

I have had the opportunity to work with a lot of wonderful colleagues and collaborators during these years. I have been enjoying the guidance and support of professor Eero Hyvönen, who has made the work possible by fertilizing my imagination and encouraging me to continue my work.

I wish to thank the pre-examiners professor Vagan Terziyan and professor Mathieu d'Aquin for valuable feedback and suggestions for improving this thesis.

When I started in SeCo, I was handed to the gentle care of my colleagues Kim Viljanen and Eetu Mäkelä. Kim has been my trusted wingman in building the ONKI service and an elevating mentor—for that I will always be grateful. Eetu has given me valuable guidance and support on methodological and technological decisions.

During my doctoral studies, I have been collaborating widely with Matias Frosterus. All the work would have been much less fun without him. I would like to thank Nina Laurenne for introducing me to the world of biological taxonomies, and diving with me in the great ball pool of colorful organisms. I wish to thank Tomi Kauppinen for his fruitful ideas on ontology services and showing by example how to be a productive researcher, and Tuukka Ruotsalo for always emphasizing the importance of scientific rigor in our work.

The ontology services presented in the thesis would be of no use without the actual ontologies. I express my gratitude to the head ontologist Katri Seppälä and Tuomas Palonen, who have been working with the ontologies

of the KOKO cloud. I wish also thank Osma Suominen, Sini Pessala, Jussi Kurki, Reetta Sinkkilä, and Robin Lindroos for participating in building the tools for supporting the ontology infrastructure, and Mikko Salonoja, Rami Aamulehto, Alex Johansson, and Henri Ylikotila for their work on the user interfaces of the ONKI service. Mikko Koho and Hannu Saarenmaa have been valuable contributors on the biological taxonomies. I am grateful for the support and collaboration of Erkki Heino, Esko Ikkala, Petri Leskinen, Arttu Oksanen, Claire Tamper, and all my other colleagues in SeCo and collaborators I have been working with during the years.

The work presented in this thesis has been carried out as part of the National Semantic Web Ontology Project in Finland (FinnONTO, 2003–2012) and Linked Data Finland project (2012–2014), both funded by the Finnish Funding Agency for Innovation (Tekes). The work has also received funding from Lusto – Finnish Forest Museum (2008), EU project MedIEQ (2006–2008), EU FP7 project SMARTMUSEUM (2008–2010), and EU FI-PPP project ENVIROFI (2011–2013). I have received grants from the Finnish Cultural Foundation (2013) and KAUTE Foundation (2015) for the doctoral research.

During the thesis work, I have also worked in the National Gazetteer of Historical Places project (2015–2016), funded by the Finnish Cultural Foundation, the Semantic Finlex project (2015–2017), funded by the Ministry of Justice and the Ministry of Finance, the Linked Open Data Science Service project (2015–2016), funded by the Ministry of Education and Culture, and the Cultures of Knowledge project (2015–2017), funded by The Andrew W. Mellon Foundation. I have received funding for short term scientific mission from the European Cooperation in Science and Technology (COST, 2016) and travel grants from the Helsinki Doctoral Programme in Computer Science (2014, 2015).

I would like to thank my parents for their support and encouragement on the path I have chosen. Finally, I wish to express my deepest gratitude to Krista for her love and understanding during the long hours.

Helsinki, May 10, 2017,

Jouni Tuominen

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 Kim Viljanen, Jouni Tuominen, and Eero Hyvönen. Ontology Libraries for Production Use: The Finnish Ontology Library Service ONKI. In *The Semantic Web: Research and Applications: 6th European Semantic Web Conference, ESWC 2009, Heraklion, Crete, Greece, May 31–June 4, 2009, Proceedings*, Lora Aroyo, Paolo Traverso, Fabio Ciravegna, Philipp Cimiano, Tom Heath, Eero Hyvönen, Riichiro Mizoguchi, Eyal Oren, Marta Sabou, and Elena Simperl (editors), Lecture Notes in Computer Science, volume 5554, pages 781–795, ISBN 978-3-642-02120-6, Springer-Verlag, June 2009.

II Jouni Tuominen, Matias Frosterus, Kim Viljanen, and Eero Hyvönen. ONKI SKOS Server for Publishing and Utilizing SKOS Vocabularies and Ontologies as Services. In *The Semantic Web: Research and Applications: 6th European Semantic Web Conference, ESWC 2009, Heraklion, Crete, Greece, May 31–June 4, 2009, Proceedings*, Lora Aroyo, Paolo Traverso, Fabio Ciravegna, Philipp Cimiano, Tom Heath, Eero Hyvönen, Riichiro Mizoguchi, Eyal Oren, Marta Sabou, and Elena Simperl (editors), Lecture Notes in Computer Science, volume 5554, pages 781–795, ISBN 978-3-642-02120-6, Springer-Verlag, June 2009.

III Jouni Tuominen, Tomi Kauppinen, Kim Viljanen, and Eero Hyvönen. Ontology-Based Query Expansion Widget for Information Retrieval. In *Proceedings of the 5th International Workshop on Scripting and Develop-*

ment for the Semantic Web at ESWC 2009, Heraklion, Greece, May 31, Sören Auer, Chris Bizer, and Gunnar Aastrand Grimnes (editors), CEUR Workshop Proceedings, volume 449, pages 52–57, ISSN 1613-0073, online CEUR-WS.org/Vol-449/ShortPaper1.pdf, May 2009.

IV Matias Frosterus, Jouni Tuominen, Sini Pessala and Eero Hyvönen. Linked Open Ontology cloud: managing a system of interlinked cross-domain light-weight ontologies. *International Journal of Metadata, Semantics and Ontologies*, 10, 3, pages 189–201, DOI 10.1504/IJMSO.2015.073879, December 2015.

V Kim Viljanen, Jouni Tuominen, Eetu Mäkelä and Eero Hyvönen. Normalized Access to Ontology Repositories. In *ICSC 2012: 2012 IEEE Sixth International Conference on Semantic Computing*, Palermo, Italy, 19-21 September 2012, pages 109–116, ISBN 978-1-4673-4433-3, IEEE, September 2012.

VI Jouni Tuominen, Nina Laurenne, and Eero Hyvönen. Biological Names and Taxonomies on the Semantic Web – Managing the Change in Scientific Conception. In *The Semantic Web: Research and Applications: 8th Extended Semantic Web Conference, ESWC 2011, Heraklion, Crete, Greece, May 29 – June 2, 2011, Proceedings, Part II*, Grigoris Antoniou, Marko Grobelnik, Elena Simperl, Bijan Parsia, Dimitris Plexousakis, Pieter De Leenheer, and Jeff Pan (editors), Lecture Notes in Computer Science, volume 6644, pages 255–269, ISBN 978-3-642-21063-1, Springer-Verlag, June 2011.

VII Nina Laurenne, Jouni Tuominen, Hannu Saarenmaa and Eero Hyvönen. Making species checklists understandable to machines – a shift from relational databases to ontologies. *Journal of Biomedical Semantics*, 5, 40, DOI 10.1186/2041-1480-5-40, September 2014.

VIII Jouni Tuominen, Nina Laurenne and Eero Hyvönen. Publishing and Using Plant Names as an Ontology Service. In *Proceedings of the first international Workshop on Semantics for Biodiversity at ESWC 2013*, Montpellier, France, May 27, Pierre Larmande, Elizabeth Arnaud, Isabelle Mougenot, Clement Jonquet, Therese Libourel, Manuel Ruiz (editors), CEUR Workshop Proceedings, volume 979, ISSN 1613-0073,

online CEUR-WS.org/Vol-979/WS_s4biodiv2013_paper_2.pdf, May 2013.

Author's Contribution

Publication I: "Ontology Libraries for Production Use: The Finnish Ontology Library Service ONKI"

The author contributed significantly to the design and specification of the ONKI library system, and the general requirements of ontology library services. The author was one of the two primary developers of the ONKI SKOS server and the ONKI selector widget, and implemented the ONKI API in the ONKI SKOS server.

Publication II: "ONKI SKOS Server for Publishing and Utilizing SKOS Vocabularies and Ontologies as Services"

The author is the first author of the publication. The author was one of the two primary designers and developers of the ONKI SKOS server and the ONKI selector widget, and primary developer of the ONKI API and SKOS support for the system. The author implemented the demonstration search interface for using ontology-based query expansion in the Kantapuu system.

Publication III: "Ontology-Based Query Expansion Widget for Information Retrieval"

The author is the first author of the publication. The author designed and implemented the query expansion functionality of the ONKI selector widget, configured the ontologies used in the use case and implemented the demonstration search interface.

Publication IV: "Linked Open Ontology cloud: managing a system of interlinked cross-domain light-weight ontologies"

The author contributed significantly to the principles guiding the building of the ontology cloud and the management process of the cloud. The author designed the publication of the KOKO ontology cloud in the ontology service, and developed the functionalities in the ONKI service to support the processing of the ontology cloud and individual domain ontologies mapped to the general upper ontology.

Publication V: "Normalized Access to Ontology Repositories"

The author participated in the design of the Normalized Ontology Repository (NOR) approach and its API. The author was the primary developer of the HTTP API used to access the ontologies in the ONKI SKOS server. The author contributed to the design of the ONKI3 user interface and the ontology metadata specification.

Publication VI: "Biological Names and Taxonomies on the Semantic Web – Managing the Change in Scientific Conception"

The author shares the first authorship of the publication. The author was the one of the two primary designers of the Taxon Meta-Ontology (TaxMeOn), responsible for the computer science expertise, and was in charge of the technical implementation of the model. The author committed the technical details related to the use cases.

Publication VII: "Making species checklists understandable to machines – a shift from relational databases to ontologies"

The author contributed to the publication equally as the first author. The author was one of the two primary designers of the taxonomic meta-ontology, and responsible for the computer science expertise, the technical comparison of the LSID and HTTP URI identifiers, and evaluation of the models for managing taxonomic information.

Publication VIII: "Publishing and Using Plant Names as an Ontology Service"

The author is the first author of the publication. The author was one of the two primary designers of the ontology model of the vernacular names, responsible for the computer science expertise, published the ontology in the ONKI service, and designed the management process of the nomenclature based on the SAHA metadata editor.

In addition to these publications, the thesis contains references to related work by the author. They include the description of the first version of the ONKI widget [268], the FinnONTO ontology infrastructure [136], the ONKI2 user interface [258], the SPARQL-based ontology service ONKI Light [248], the deployment and further development of the ONKI service by the National Library of Finland [249], and the use of ontology services in the legal domain [87].

The author has also published other work related to ontology services, including using the ONKI service in cultural heritage [133], birdwatching [129], for publishing historical places as on ontology time series [135], and for visualizing automatically enriched ontology relationships based on co-occurrences of concepts in annotations [153]. Furthermore, the author has worked on publishing ontologies in a linked data service [134] and developing a federated ontology service for historical places [130].

1. Introduction

1.1 Background and Research Environment

As the amount of information in information systems grows, it is harder for the users to find relevant information for their needs [137]. Not only is the information hard to find, but it is not connected to other relevant information—thus getting an extensive understanding about a topic is challenging. These issues are intensified in the massive World Wide Web, which was estimated to contain 11.5 billion indexable web pages already in 2005 [102], and almost 50 billion of them in a more recent study [263].

The full text search employed by many search engines has several limitations. A simple search algorithm does not distinguish significant words from non-significant ones in the document, leading to the loss of precision of the search [46]. The issue can be compensated by ordering the search results based on their assessed relevancy for the search task [181], effectively displaying the most relevant results first. On the other hand, all significant terms regarding the information content of the document might not appear in the document, decreasing the recall of the search [114].

The Semantic Web[1] [28, 77] is an extension of the current World Wide Web, providing technologies for processing information based on their semantics. Managing textual contents on the conceptual level resolves the issues of purely lexical full text search, such as handling synonymy and homonymy. A practical subtopic of the Semantic Web is the Linked Data [27, 112] concept, which is a method for publishing data in an interlinked way. The semantic interoperability and interlinking of the information contents in the web changes the nature of the web from the web of documents to the web of data. These technologies enable building of

[1] http://www.w3.org/2001/sw/

services on top of the integrated datasets, for example, providing users novel search and recommendation interfaces, thus improving information findability.

Ontologies are at the core of the Semantic Web infrastructure, as they model the domains of interest in a formal, machine-understandable way. An ontology acts as a shared conceptualization of a domain [98, 35, 244], enabling different parties to use a common language when communicating about the domain [100]. When information objects are described with metadata [222, 17, 71, 46] referring to concepts of an ontology, machines can interpret their meanings. This semantic content annotation enables, e.g., the integration of heterogeneous data collections and automatic reasoning based on the properties of concepts [143, 239].

An ontology can be built by a domain expert, but also methods for automatic and semi-automatic ontology generation exist [176]. Also the content annotation can be done manually or (semi-)automatically [159, 234, 276]. In addition to ontologies, semantically lighter knowledge organization systems (KOS) originating from library and information science, such as subject headings, classifications, and thesauri can be used to harmonize the used terminology in content descriptions and search [124, 95]. KOSs can also be utilized in, e.g., query expansion, cross-language search, and as a navigation aid for accessing contents. Knowledge organization systems can be interlinked in order to facilitate the integration of data described using different KOSs, by utilizing the methods of ontology mapping [127] and matching [233].

An alternative method for using explicit ontology-based metadata for improving the information findability is to use automatic methods analyzing the contents of information objects. For example, natural language processing methods can be used to identify the meanings of words based on their context in a text document [53]. However, the strength of using explicit metadata is its applicability also to non-textual objects, such as images and videos, as otherwise text search would not be possible without reliable content analysis methods.

For facilitating the use of ontologies, specialized software systems—ontology servers, have been proposed for publishing ontologies and providing services for using them [79, 68, 9, 108, 60]. Most of the ontology server implementations introduced in the Semantic Web research have been designed for developing ontologies, and not for their actual usage, e.g., in content annotation or information retrieval, and therefore lack

crucial functionalities needed in applications [9]. The common functionalities of ontology servers include user interfaces for visualization and browsing of ontologies, and searching for concepts in an ontology. Several implementations also provide application programming interfaces (API) for the programmatic use of the ontologies. In addition to APIs, ontology functionalities may be integrated into client systems with user interface components.

1.2 Objectives and Scope

The aim of this thesis is to provide methods and technological solutions for publishing knowledge organization systems in such a way that they can be utilized cost-effectively in external applications. As a solution, the notion of an ontology service is presented. An ontology service is a software system that can be used by ontology developers for publishing their ontologies, and by content indexers, information searchers, and application developers to use ontologies in their tasks. The user needs for knowledge organization systems are analyzed, and based on them a set of requirements for functionalities is formulated. As a proof-of-concept system, implementations of such an ontology service are provided and their application to real life cases is reported. The KOSs used in the cases include Finnish and international thesauri, lightweight ontologies covering general and domain-specific concepts, and semantically richer biological nomenclatures and classifications.

The objectives of the ontology services presented in this thesis are:

- **Ontology publication channel.** Provide a complete publication workflow for the ontology developers to publish an ontology, or a new version of it.
- **Heterogeneous ontologies.** Support for distinct ontology formats by using a harmonizing data model and configuration options.
- **Tools for metadata creation.** Provide means to ontology-based content indexing.
- **Support for distributed content creation.** Facilitate content creation in distributed workflows, where content is curated by independent parties and aggregated into one system, e.g., a web portal.
- **Facilitate search tasks.** Support the use of published ontologies in

information retrieval by offering functionalities, e.g., for query expansion.

- **Multiple ontologies and repositories.** The users should be able to access multiple ontologies, even originating from different ontology services, simultaneously in a coherent way.
- **Programmatic access.** Applications should be able to use the ontologies via application programming interfaces (API) for searching for concepts and getting their properties.
- **Evaluation by applying into practice.** The applicability of the services will be tested by building a proof-of-concept system, which is piloted in real life scenarios.
- **Promote complex KOSs.** The system should support not only simple, but also richer knowledge organization systems.

Based on these objectives that guide the design and implementation of the ontology services this thesis seeks to find solutions for the following research questions (RQ):

1. How can lightweight ontologies be published on the Semantic Web so that they can be utilized in content indexing and information retrieval tasks?

2. How can a collection of independent or interconnected ontologies—in different formats and repositories—be published and utilized using shared user interfaces and APIs?

3. How can richer knowledge organization systems, such as biological nomenclatures and classifications, be managed as an ontology and published using an ontology service?

The research questions are answered with the publications I–VIII. Table 1.1 shows which research questions the individual publications contribute to. The contributions of the publications are summarized in Chapter 3.

1.3 Research Process and Dissertation Structure

The research presented in this thesis has been conducted by applying the methodologies of design science [182, 116, 211] and action research [24, 42,

Research question	PI	PII	PIII	PIV	PV	PVI	PVII	PVIII
RQ1	x	x	x					
RQ2				x	x			
RQ3						x	x	x

Table 1.1. The relationship between the research questions and the publications.

62].

Design science is a technology-oriented paradigm in the information systems discipline that aims to create things that serve human purposes [182]. The significance of a research is determined by the value or utility it provides—does it work, is it an improvement? Instead of new theories, the outcomes of design science are innovative and useful artifacts, which include constructs, models, methods, and implementations. The process of design science includes two phases: building and evaluation. The nature of design science tends to be applied: it exploits knowledge created by basic research to develop new technologies. However, the created artifact and its working environment might not be well understood, and in such case the artifact itself presents new scientific questions. By designing, building, and applying an artifact, knowledge and understanding about a problem domain and its solution is achieved [116]. As opposed to routine design and systems building work, design science builds novel ways to solve important, unsolved problems or provides more effective or efficient ways to address previously solved problems. The solutions are generalizable and provide new knowledge for the application domain. The applicability of the artifact is evaluated in real-world scenarios by observational, analytical, experimental, testing, or descriptive methods.

Complementing the technological aspect of design science, action research emphasizes the social elements of information systems research. In an action research setting, scientists and the subjects of the study collaborate in order to study and solve problems in organizations [24]. The research involves two phases: the diagnostic stage to analyze the current situation and the therapeutic stage to carry out changes to improve the situation. In contrast to case studies, in action research the researcher is involved in the studied phenomenon and the research is carried out in a more rigorous way [23]. The rigor is ensured by following the action research cycle: diagnosis, action planning, action taking, evaluating, and specifying learning.

The design of the ontology services presented in this thesis is based on

analyzing the user requirements of ontology users and existing ontology server implementations by a) conducting a literature and system review, b) formulating illustrative use scenarios, and c) running a prototype system as a living lab service, gathering feedback from the actual users of the system. Based on the cyclic nature of design science and action research, similar methods have been used to evaluate the purposefulness of the developed ontology services. The prototype system itself acts as a proof of concept, demonstrating the utility or suitability of the software artifact for the given requirements [210]. Furthermore, using the system in an action research setting in real use cases evaluates the effects of the system use in real-world situations. By basing the functionalities of the system on existing research and illustrative use scenarios, the utility of the system is ensured.

This thesis is organized as follows. The theoretical background of the work is presented in Chapter 2. Building on the theory and based on the publications included, the results of the thesis are summarized in Chapter 3. Finally, the implications of the results, the validity of the work, and further research are discussed in Chapter 4.

2. Theoretical Foundation

2.1 Modeling Knowledge Organization Systems

2.1.1 Knowledge Organization Systems

Knowledge organization systems (KOS) originate from library and information science, where they are used as schemes for organizing information and promoting knowledge management [124, 95]. Examples of different types of KOSs include classification schemes, subject headings, authority files, taxonomies, thesauri, and ontologies [124, 119, 238, 95]. They provide a controlled vocabulary in the given domain of interest, and harmonize the terminology used to describe the information items in information collections, e.g., in digital libraries or document databases.

In its simplest form, a controlled vocabulary is a list of terms, where each term corresponds to a concept of the domain. It can also include other information about the terms, such as synonyms, descriptions, and source information. A taxonomy arranges the terms in a controlled vocabulary into a hierarchy, aiding the users selecting a suitable term in, e.g., content description or information retrieval [91]. Extending taxonomies, thesauri may include richer information about the terms, such as associative relations between them [91]. Guidelines for creating, displaying, and managing thesauri are documented in international and national standards, such as ISO 25964 [6, 66] and SFS 5471 [2].

Thesauri and other controlled vocabularies are used primarily for improving information retrieval [11, 236]. This is accomplished by using the concepts or terms of a thesaurus in content indexing, content searching, or in both of them, thus simplifying the matching of query terms and the indexed resources (e.g., documents) when compared with using free,

uncontrolled natural language. The relations of thesauri can be utilized in information retrieval, for example by expanding query terms to more specific terms based on the concept hierarchy. Multilingual thesauri can be used for cross-language search where the information contents or the related metadata are expressed in different language than the one used by the end user.

Knowledge Organization Systems, such as thesauri, are of great benefit for the Semantic Web [19, 282, 193, 106, 262, 278], enabling semantically disambiguated data exchange and integration of data from different sources, though not to the same extent as ontologies [240] where the semantics of concepts is defined in more refined and machine-understandable ways [232, 10]. Ontologies based on thesaurus-like structures can be called lightweight ontologies [82, 159, 93, 143].

The Simple Knowledge Organization System (SKOS) [188, 19, 189] developed within W3C is a data model and a syntax for expressing concept schemes such as thesauri, and is largely compatible with the ISO 25964 thesaurus standard [66, 139]. SKOS provides a standard way for creating vocabularies and migrating existing vocabularies to the Semantic Web. SKOS solves the problem of diverse, non-interoperable thesaurus representation formats by offering a standard convention for presentation. Existing thesauri can be transformed into SKOS format via conversion processes [261, 245, 193, 237, 282]. When a thesaurus is expressed as a SKOS vocabulary, it can be processed with standard RDF/SKOS tools in a uniform way.

There are also methods for converting thesauri into semantically richer OWL ontologies [262, 136, 44, 162]. Compared with SKOS conversion techniques, the OWL-based methods cannot be fully automated, as they require human effort for refining the semantic relations of the concepts [162]. Especially the *is-a* hierarchy of the ontology needs to be carefully constructed since the hierarchy of a thesaurus may have been built using a mix of different hierarchical relations [136, 162], which cannot be used as is for, e.g., subclass reasoning. The use of existing thesauri as the basis for ontologies enables the backwards compatibility with legacy data annotated with the thesauri, and facilitates the publication of the data as Linked Data.

This thesis seeks to develop publication methods for the cost-effective utilization of KOSs in, e.g., content indexing and information retrieval. In this context, SKOS is applied as a harmonizing model for representing

KOSs.

2.1.2 Interlinked Ontologies

In Linked Data paradigm, entities can be linked on many levels: data instances [112, 140, 103], metadata schema fields [267], and concepts of ontologies [286] can be interlinked to facilitate interoperability between datasets. Linking the data through ontologies allows additional interoperability due to the inferred knowledge gained through the shared ontology semantics [121]. When integrating datasets that use different ontologies (or KOSs), the ontologies need to be reconciled. Ontology reconciliation [104, 283] is a broad term, covering ontology merging, alignment, and integration. Most of the reconciliation methods are automatic or semi-automatic, which can lead to lower quality [32], especially if the ontologies were originally expert-made [73].

To faciliate the interoperability between different ontologies, there have been efforts to establish guidelines for the creation and management of ontologies, e.g., in the context of the OBO Foundry initiative [235]. The focus in OBO Foundry is on coordinating the development of different ontologies in the biomedical domain under shared principles. General, domain-independent ontology design principles have been proposed by several researchers [99, 260, 101, 191, 273, 89]. Linked Open Vocabularies (LOV) [267] is an effort on building a high quality catalogue of reusable vocabularies, and making their interconnections visible. Instead of KOSs, LOV focuses on metadata schemas.

Ontology linking methods are used also in ontology modularization [243, 7], where ontologies are divided into smaller interlinked parts to facilitate distributed development and re-use. There have been several efforts on building a general upper ontology [184] that can be used as a foundational basis for domain ontologies. Some of the upper ontologies have been developed from scratch, such as CYC [170], while, e.g., the Suggested Upper Merged Ontology SUMO [194] was created by merging existing ontologies.

For ensuring the consistency of interlinked ontologies, the changes of the ontologies have to be communicated to the dependent ontologies, e.g., by applying methods from the field of ontology evolution [281, 111, 169, 81, 128]. Methods include the detection of changes in an updated ontology by using logs [242, 157, 144] or comparing two versions of the ontology [164, 272]. To facilitate the processing of changes, different change types can be

identified [246, 186], or more abstract change patterns can be constructed from atomic changes [160, 157, 144]. There has also been research on the nature of the change types the users are most interested in [37]. Also, the extra challenges of distributed ontology development [160, 241, 175, 146] have to be addressed when operating with interlinked ontologies.

This thesis aims to design methods and tools for managing and publishing an interlinked cloud of cross-domain ontologies in such a way that they can be utilized using shared user interfaces and APIs. The approach is based on modularizing the ontology development work into an upper ontology and domain ontologies extending it, and keeping track of their semantic dependencies.

2.1.3 Biological Nomenclatures and Taxonomies

Management of names and taxonomies of organisms in biology is an example use case for the need of a complex KOS that cannot be represented using simple, general KOS presentation languages, such as SKOS. In biology, taxonomy refers to the discipline that identifies, describes, and names groups of organisms (taxa) based on their shared characters [150]. The organism groups are organized into taxonomic hierarchies. Taxon names and classifications are important when integrating biological data from multiple sources [225, 201, 145, 253], and are therefore considered central resources, for example in biodiversity management [25, 200, 228, 215, 85, 86, 105, 206, 219]. The changing nature of the names poses challenges for their management [148, 166, 208, 229].

There are several issues that make the biological nomenclatures and classifications a suitable domain for the study of the modeling and publishing a rich KOS as an ontology. 1) Biological names are not stable or reliable identifiers for organisms as they or their meaning change in time. 2) The same name can be used by different authors to refer to different taxa, and a taxon can have more than one name. 3) Taxonomic knowledge is changing all the time and increases due to new research results. The number of new organism names in biology increases by 25,000 every year as new taxa to science are discovered [163]. At the same time, the rate of changes in existing names has accelerated by the implementation of molecular methods suggesting new positions to organisms in taxonomies. 4) The notion of 'species' in the general case is actually very hard to define precisely. For example, some authors discuss as many as 22 different ways of defining the concept of species [185].

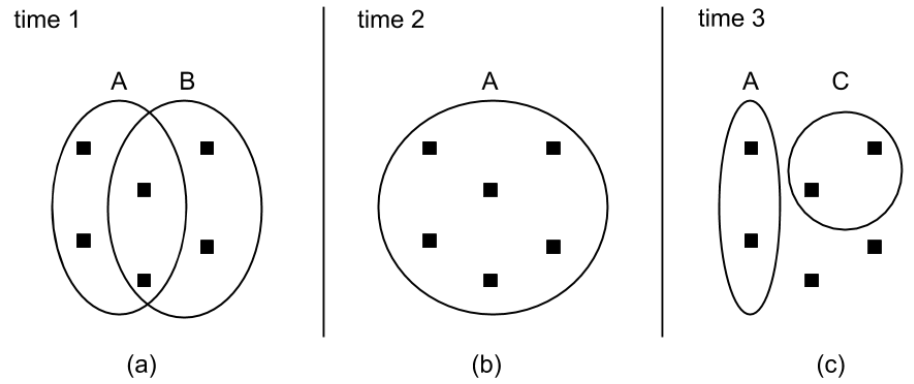


Figure 2.1. A hypothetical example of changes in taxonomic concepts and taxon names in time. First, two separate species A and B partly overlap (a). Then the two species are merged into a single species that has the name A, and the name B becomes a synonym to the name A (b). Finally, the species A is split into two separate species. The name A remains a valid name with a narrower meaning, and the name C is given to the new taxonomic concept. The black squares illustrate the biological characters of organisms, and the ellipses describe the limits of taxonomic concepts.

Although biological naming convention is regulated by nomenclature codes, e.g., International Code of Zoological Nomenclature (ICZN) [138] and International Code of Nomenclature for algae, fungi, and plants (ICN) [187], the names cannot be used as reliable identifiers when referring to taxa due to their ambiguity. Figure 2.1 depicts a typical series of changes in taxonomic concepts and their names. The border between the two species is unclear, and later the two species are merged into a single one, which is finally split into two (or more) species again. The taxon names may or may not change accordingly. Both the taxonomic concepts and their names change over time, and tracing the meaning of a name is impossible without a reference to a study. The reason for the continual changes is that every study has a different set of taxa, biological characters, and methods, and consequently their results are different.

A species checklist is a collection of names of organisms of a certain taxonomic group, compiled into a single taxonomic hierarchy by scientific experts. Checklists often present the species occurring in a particular geographical area. Comprehensive reference lists and catalogues of the names have been proposed as a solution to facilitate the access to the names and harmonize their usage [285, 105, 65, 208, 225, 173]. The need for such a list has been recognised, e.g., for vascular plants by the Convention on Biological Diversity (CBD) [5].

There are efforts on curating or aggregating taxon names from authoritative sources covering all species groups in the world, such as the Catalogue of Life (CoL) [223], the Encyclopedia of Life (EoL) [207], the Universal

Biological Indexer and Organizer (uBio) NameBank [225], WikiSpecies[1], the NCBI Taxonomy database [76], Open Tree Taxonomy [120], GBIF ChecklistBank [94], Index to Organism Names (ION)[2], and BioNames [203]. Other efforts are focusing on specific taxonomic groups or regions, such as ZooBank [215], the International Plant Names Index (IPNI) [55], World Register of Marine Species [51], Fauna Europaea [64], and the Atlas of Living Australia[3].

Berendsohn [25] introduced the concept of a "potential taxon" to overcome the name ambiguity issues. A potential taxon is a combination of a taxon name and a literature reference to the taxonomic concept, that can be used, e.g., in databases for taxon references, enabling the interlinking of differing taxonomic views [26]. The importance of persistent identifiers for organism names has been further discussed by several researchers [200, 209, 225, 205, 219].

Pullan et al. [214] presented the Prometheus data model for managing taxonomic nomenclature and multiple related classifications separately, while Ytow et al. have developed the Nomencurator data model for representing and managing taxonomic nomenclature in a relational database [280]. Page [200] presented a simple data model for presenting taxon names and their relations, using using Life Science Identifiers (LSID) for identifying them. The use of LSIDs has been suggested also by organizations publishing taxonomic data [212, 56, 221], and piloted, for example in the Catalogue of Life database [148]. Further, Schulz et al. [228] presented an ontology model of biological taxa and its application to physical individuals. The model is based on a single unchangeable classification. Franz and Peet [85] formulated the use of semantics in relating taxa to each other, within a single taxonomic hierarchy and between two distinct hierarchies. Franz and Thau [86] evaluated the limitations of applying ontologies to the scientific names and concluded that ontologies should focus either on a nomenclatural point of view or on strategies for aligning multiple taxonomies.

As a use case for taxonomic ontologies, Lepage et al. [171] have implemented the Avibase database system for managing and organizing taxonomic concepts from major bird taxonomic checklists. There are also practical efforts on publishing taxonomic concepts as Linked Data, such

[1] `http://species.wikimedia.org`
[2] `http://www.organismnames.com`
[3] `http://www.ala.org.au`

as Taxonconcept.org[4] and Geospecies[5] that aim to provide Linked Open Data identifiers for species concepts and link them to related data from different sources. Chawuthai et al. [48] have presented an ontology model for managing the change of taxonomic concepts and publishing them as Linked Data.

In addition to data models that are focused on taxonomic information, there are metadata schemas for exchanging biodiversity data on broader scope, such as Darwin Core (DwC) [61, 277], the related Taxonomic Concept Transfer Schema (TCS) [251, 155], created by the Biodiversity Information Standards (TDWG), Access to Biological Collection Data (ABCD) [125], and Biological Collections Ontology (BCO) [271]. Darwin Core has a set of taxonomic extensions, Global Names Architecture (GNA) Profile [220], that introduces properties for richer nomenclatural details of taxa. Also, a semantically refined version of Darwin Core, Darwin-SW [22], has been proposed.

Methods of ontology versioning [161], evolution [195], and matching [233, 74] are relevant in management of taxonomic and nomenclatural information, as there exist multiple views on taxonomy and taxonomic knowledge changes as new research results are published. Thus, systems that support the usage of multiple versions of ontologies [128] and concepts [252] simultaneously are needed. There are approaches that are focused on life science ontologies, providing support for mapping ontologies [97, 110, 158], and systems for matching taxonomic concepts between databases [115, 202, 148, 51, 36, 218, 266, 84].

This thesis uses biological nomenclatures and classifications as an example use case of modeling and managing a richer KOS as an ontology. The focus is on practical management of the names and their changes.

2.2 Publishing and Using Knowledge Organization Systems

2.2.1 Ontology Servers

Once an ontology is modeled and serialized in some format, such as SKOS or OWL, it can be published for the wider community to be used as a shared domain model. In order to facilitate the usage of ontologies, on-

[4] http://www.taxonconcept.org
[5] http://lod.geospecies.org

tology servers [79, 68, 9, 108, 60] have been proposed for publishing ontologies and vocabularies on the web. Together with ontologies they have been considered a key resource for enabling the vision of the Semantic Web [136, 18, 108, 60]. The motivation for publishing ontologies using ontology servers instead of making them available as mere data files is to support the use of ontologies in applications. Ontology servers can provide ready-to-use services that can be integrated into information systems in a cost-effective way. Without such services, the user organizations have to implement the common functionalities for accessing ontologies in their own systems, leading to redundant work.

Parallel terms with ontology server are ontology library, ontology repository, and ontology service, each with a slightly different emphasis on the topic. In addition, the community of Networked Knowledge Organization Systems (NKOS)[6], that aims to develop web-based information services to support the description and retrieval of diverse information resources using KOS, uses the terms terminology registry and terminology services [95]. Common to these systems is that they are intended for publishing, managing, sharing, finding, and reusing ontologies and vocabularies for content indexing, information retrieval, content integration, and other purposes. Traditionally, the main focus in ontology server systems has been in supporting ontology development instead of the runtime usage of ontologies such as indexing and ontology-based end-user applications [68, 9]. The features of the systems vary greatly as they are designed for different purposes and based on specific user requirements [60].

An ontology server can support different phases in the ontology lifecycle, which can be defined as 1) design, 2) commit, and 3) runtime [9], or in a more fine-grained way as 1) acquisition, creation, and modification of vocabularies, 2) publication of vocabularies, 3) access, search, and discovery, 4) use, and 5) archiving and preservation of vocabularies [95]. The different lifecycle phases involve different user groups, which can be classified into ontology developers, ontology users, and application developers [60], or similarly in the NKOS community into KOS owners or creators, end users, and system developers [95]. The design phase covers tasks involved in ontology development, such as ontology engineering or editing, storing, versioning, mapping, and publishing. Once an ontology is published, the commit phase refers to the activity where a user is trying to find a suitable ontology for her needs, and needs support for discovering and evaluating

[6] `http://nkos.slis.kent.edu`

candidate ontologies. In the runtime phase, the user needs tooling for finding concepts for a given task, such as content indexing. Based on a survey of ontology servers, d'Aquin and Noy [60] have identified the key functions of an ontology server to be search, browsing, selecting and evaluating ontologies, and programmatic access to ontologies. Similar features have been recognized also by other researchers [108, 18, 95].

Most ontology servers are web-based systems that catalogue ontologies on a specific domain, such as biomedical sciences [235, 279, 196, 52], agriculture[7], oceanography [224, 167], government[8], by a specific organization, such as Library of Congress Linked Data Service[9], or with no such restriction. They provide access mechanisms to ontologies as user interfaces, APIs, or both of them [9, 60]. The user interfaces typically include search and browsing functionalities for finding ontologies and concepts in them. Search functionalities can be provided as string search based on the labels or other textual properties of the concepts, and utilizing the semantic relations of the concepts. Browsing interfaces typically visualize the structure of an ontology as a hierarchical tree or as a graph, where the concepts are presented as nodes and the relations as arcs between them. Ontology servers can provide users with listings of available ontologies, which are often classified based on different criteria. In some implementations, the set of ontologies can be further filtered and investigated using a faceted search.

Some ontology servers have been implemented for publishing a single, specific ontology, such as SUMO Browser [230] and GTAA Browser [40], whereas some systems focus on providing a directory-like listings of ontologies, such as DAML Ontology Library[10], Protégé Ontology Library[11], OBO Foundry [235], oeGOV, OntologyDesignPatterns.org [33], SHOE ontology library [113], Basel Register of Thesauri, Ontologies & Classifications (BARTOC) [168], and TaxoBank [122]. There are systems that focus on the ontology development and editing functionalities, such as iQvoc [20], PoolParty [226], VocBench [43], TemaTres[12], SKOS Shuttle[13], TopBraid Enterprise Vocabulary Net [255], SKOS editor [50],

[7] http://agroportal.lirmm.fr
[8] http://oegov.us
[9] http://id.loc.gov
[10] http://www.daml.org/ontologies/
[11] http://protegewiki.stanford.edu/wiki/Protege_Ontology_Library
[12] http://www.vocabularyserver.com
[13] http://skosshuttle.ch

PeriodO gazetteer [231], Neologism [21], Open Metadata Registry [213], WebOnto [69], Adapted Ontology Server [49], Ontolingua Server [75], and Medical Ontology Server [92].

Inter-ontology relations are considered important in several ontology servers, such as BioPortal [196], ACOS [172], MMI Ontology Registry and Repository [224], CATCH Vocabulary and alignment repository [264], ROMULUS [156], and Ontohub [190], as they support the creation and representation of concept mappings. There are systems that emphasize community-based aspects, such as uploading, rating and commenting on the ontologies. These ontology servers include, e.g., BioPortal, ACOS, and CupBoard [58].

There exist several search engines that crawl the web for RDF data and index them, such as Swoogle [67], Watson [59], and Sindice [198] for any data, whereas OntoSelect [41] and Falcons [217] are designed especially for ontologies. OntoSearch2 [254] is a similar ontology search engine, but it uses its own repository as opposed to crawling the public web. Such systems can be useful when searching for suitable ontologies to use in applications, and provide an overview of web-published ontologies in a specific domain.

Many ontology servers provide APIs for accessing the ontologies, typically for querying for ontologies and/or their concepts, and getting information about them. There are several specifications for APIs, such as Foundation for Intelligent Physical Agents (FIPA) Ontology Service [4] and Ontology Web Services (OWS) [57]. There are API implementations for processing ontologies in programming languages, e.g., the Java-based SKOS API [151], OWL API[14], and the APIs in DOGMA Server [141] and KAON Server [197]. For accessing ontologies on the web, there are several ontology servers that provide Web Service (SOAP) APIs, e.g., SKOS Web Service API [257], OCLC Terminology Services [269], Ontology Lookup Service (OLS) [52], Watson, CATCH Vocabulary and alignment repository, and NERC Vocabulary Server [167]. A more recent approach is to provide access to ontologies through a RESTful HTTP API, such as in the case of SISSVoc [54], OCLC Terminology Services, Otago Ontology Repository [204], BioPortal, iQvoc, PoolParty, NERC Vocabulary Server, HIVE [96], TemaTres, and SKOS Shuttle. SPARQL [107] is the standard way to provide an application interface to Semantic Web databases, and it can be used also to access ontology repositories. Similarly, general RDF

[14] `http://owlapi.sourceforge.net`

libraries can be used for processing ontologies, such as Apache Jena[15] and RDFLib[16]. Ontology servers such as OCLC Terminology Services and BioPortal provide widgets, user interface components that can be integrated into applications, to enable ontology-based search functionalities in applications.

To facilitate simultaneous access to several ontology servers, shared access mechanisms and protocols are needed. Open Ontology Repository (OOR) [18] is an initiative that aims to specify an architecture and interfaces for interoperability between ontology repositories. Also other researchers have emphasized the importance of interoperability between ontology repositories [60, 108]. Ontohub is an ontology repository engine that follows the ideas of the OOR initiative, and provides inter-repository access by defining a generalized federation API that needs to be implemented in the participating repository or as a wrapper around the legacy API of the repository. Common Ontology API Tasks (OntoCAT) [8] is a programming interface to query multiple ontology repositories seamlessly from an application. The system is based on wrappers that are implemented for each supported ontology repository. OLS2OWL [90] is a plugin for Protégé ontology editor enabling simultaneous queries to multiple ontology servers by using a similar wrapper approach as in OntoCAT. JSKOS-API [168] is a general HTTP API for accessing knowledge organization systems, with methods for concept search and lookup. By implementing the API in multiple ontology servers or using wrappers, it is possible to provide inter-repository search and browsing interfaces. There are also general protocols for accessing knowledge bases, such as the Open Knowledge Base Connectivity (OKBC) [47], and agent communications languages FIPA-ACL [3], the Knowledge Query and Manipulation Language (KQML) [80], and the Semantic Agent Programming Language (S-APL) [152].

Regarding the different aspects of ontology servers, the focus of this thesis is to provide publication channels for ontologies in different formats and support for their runtime use, e.g., in a network of distributed content creation. One of the design principles is the support for the use of multiple ontologies and ontology repositories simultaneously.

[15] `https://jena.apache.org`

[16] `https://rdflib.readthedocs.io/en/stable/`

2.2.2 Semantic Annotation and Information Retrieval

The typical use cases for ontologies and other knowledge organization systems are semantic annotation and information retrieval. Both these tasks can be facilitated with ontology services. In ontology-based manual annotation human cataloguers create content descriptions using ontological concepts [159, 199, 16]. The annotation process is usually guided by indexing guidelines and conventions that may be general [1, 15] or shared by particular disciplines or organizations. The created metadata can be based on a specific schema, which can be a simple collection of key-value-pairs, such as Dublin Core [63], or a semantically richer model with relations between the individual metadata fields [250, 227]. The annotation task can be defined as a process of analyzing the content to be annotated, finding relevant ontological concepts from the selected ontologies, and storing them in metadata fields. The process can be streamlined and made more effective with different kinds of automated tools [17].

Based on a user study, Hildebrand et al. [118] have identified the following use cases of a human annotator for a concept search:

- The user already knows the concept she would like to use as a descriptor for the content, and wants to find the concept from the used thesauri.
- The user does not know the most suitable concept for the content description beforehand and she needs to examine the thesauri to find one.
- The user suspects that the used thesauri do not contain a concept she would need for the task, and she needs to ensure this before she adds a new concept into one of the thesauri.

The human annotator's task to find the best matching concepts for her needs can be aided by providing concept search and browsing functionalities. These general functionalities can be provided by ontology servers as APIs and user interface components that can be used for integrating them into applications. The importance of such services for thesaurus use, sharing, and interoperability has been emphasized by several researchers [95, 183, 216, 284, 30, 119, 124]. Such an approach relates to the notion of service-oriented architecture (SOA) [231, 70, 154], where software components are provided as technology independent services to other applications to be used over a network through a communication protocol.

Regarding the ontology services discussed in the previous section, OCLC Terminology Services provides a widget for querying controlled vocabularies and displaying information about their terms in the sidebar of the Internet Explorer browser, and transferring selected terms into a web-based cataloging application [269]. BioPortal provides web widgets that can be integrated into web applications, for concept search, selection, and visualizing ontologies [196]. The use of SKOS Web Service API as web widgets has been demonstrated in the STAR project by providing functionalities for concept search, expansion, and presentation [31].

Hildebrand et al. [117] have implemented a configurable autocompletion search widget for RDF repositories, based on which Amin et al. [14] have conducted a user study on the organization strategies for the autocompletion suggestions to provide effective means of navigation and finding relevant terms. Further, Hildebrand et al. [118] performed a user study in which museum professionals used the widget for annotation. They reported on different strategies for matching, sorting, grouping, and displaying contextual information for the autocompletion suggestions. Malaisé et al. [179] conducted a user study on expert annotators using GTAA Browser, and concluded that the users mainly used the alphabetical search functionality to find relevant concepts, whereas the concept hierarchy browser was not used that much.

The ability of the information retrieval systems to produce relevant search results depends on the user's ability to represent her information needs in a query [270]. If the vocabularies used by the user and the system are not shared, or if the vocabulary is used in different levels of specificity, the search results are usually poor. Query expansion has been proposed to solve these issues and to improve information retrieval by expanding the query with terms related to the original query terms [45]. Query expansion can be based on a corpus, e.g., analyzing the co-occurrences of terms, or on knowledge models, such as thesauri [83, 274] or ontologies [270]. Methods based on knowledge models are especially useful in cases of short, incomplete query expressions with few terms found in the search index [270, 274].

Ontology-based query expansion can be used interactively to guide the user to formulate his query, for example by providing an autocompletion text search for disambiguating and selecting ontological concepts [132], and automatically by adding concepts to the initial query based on their ontological relations [34, 123, 29, 180, 149]. Typical relationships used in

query expansion are the synonym, hyponym, hypernym, and associative relations [14, 126, 147, 142, 72]. When considering general associative relations, caution should be exercised as their use in query expansion can lead to an uncontrolled expansion of result sets, and thus to potential loss in precision [256, 126]. In an experiment by Navigli and Velardi [192], it was noted that extracting the query expansion terms from the sense definitions of the query terms from an ontology produced better results than using the taxonomic relations (e.g., synonym, hypernym). Ontology-based query expansion can also be used for cross-language information retrieval [45], and in addition to general-purpose ontologies, domain-specific ontologies can be utilized, e.g., for spatial query expansion [88]. A concrete example of a terminology service with query expansion support is the FACET Web demonstrator [30] that provides a web service and user interface for accessing thesauri.

This thesis aims to develop practical solutions and tools for ontology-based content annotation and information retrieval. The common user tasks of such workflows are catered by providing user interface components and APIs that can be integrated into external applications.

3. Results

In the following, the results to the research questions of this thesis are presented. Furthermore, the results are reflected against previous research in Chapter 4.

3.1 Ontology Services (RQ1)

The research question 1 concerns publishing thesauri in such a way that they can be easily and cost-effectively used in ontology-based workflows, especially in content indexing and information retrieval.

1. How can lightweight ontologies be published on the Semantic Web so that they can be utilized in content indexing and information retrieval tasks?

The publications I–III provide solutions for this question by presenting ontology services. The main idea is to publish the ontologies not only as data, but as services that can be used by humans and machines for integrating ontology-based functionalities into applications, provided as user interfaces and application programming interfaces (API). The functionalities of the developed ontology services were developed based on the analysis of the user groups of ontologies, acting in different phases of the life cycle of an ontology.

The main user groups of ontologies were identified as 1) ontology developers, 2) content creators, 3) information searchers, and 4) software developers. In Publication I, their needs for using ontologies are classified into the following tasks.

1. **Designing the ontologies.** The structure and modeling principles of an ontology are developed based on the analysis of the subject domain

and the use cases of the ontology. Ontology developers need tools for creating the ontology, collaborative editing, reuse, and alignment.

2. **Populating the ontologies.** An ontology may contain a high amount of instances, e.g., people, organizations, and places. To save efforts, they can be harvested from existing sources, or collected from the end users. Tools may provide support for content collection and updating processes.
3. **Publishing the ontologies.** To promote the use and reuse of an ontology, the ontology owners can make it accessible to the users. Prior to publishing the ontology, its quality can be ensured with manual and automatic methods.
4. **Finding, comparing, and committing to ontologies.** When there is a need to use an ontology in an information system, the information architect of the system needs support for selecting a suitable ontology for her needs.
5. **Ontology-based semantic application creation.** Application developers need to learn, evaluate, and apply methods for integrating ontologies into applications, e.g., by using user interface components and APIs.
6. **Ontology-based semantic content creation.** The work of content indexers can be facilitated by providing tools, e.g., for browsing the ontologies, searching and selecting concepts, and (semi-)automatic indexing.
7. **Ontology-based end-user applications.** Application developers can utilize an ontology for building end-user applications, such as semantic portals, to facilitate information findability. The ontology can function as an educational resource for end users to learn about its domain.

The common functionalities for utilizing ontologies in ontology-based applications of different kinds were identified as concept search, browsing, and selection. The developed user interfaces, widgets, and APIs are designed to support these basic tasks. The proposed ONKI ontology service is based on several implementations of ontology servers suited for ontologies of different kinds due to distinct needs for accessing them. For example, a natural way to display a thesaurus is a tree-like visualization, whereas the users of geographical ontologies may prefer map-based user interfaces. Publication II presents an ontology server for thesaurus-like, lightweight ontologies, the ONKI SKOS system. The system supports publishing of

syntactically, semantically, and structurally differing thesaurus formats, with the requirement that the thesaurus has to be in RDF format and have some basic structure including concepts, their labels, and possible inter-concept relations.

Existing thesauri in legacy formats can be converted into W3C's SKOS data model, which acts as a harmonizing model for expressing knowledge organization systems. Such thesauri can be published in the ONKI SKOS server, which then provides functionalities for the users of the ontology. Thus, organizations developing thesauri do not have to implement their own thesaurus-specific publication systems and the users can use different thesauri with shared tools. The real-life benefits of using the ONKI SKOS server have been demonstrated by applying the system to use cases of a) content indexing in health promotion and cultural heritage context, among others, and b) information retrieval in the collections of the Finnish museums of forestry. The ONKI service enables content creation in a distributed network of organizations, where each organization uses shared ontologies and ontology services for accessing them, thus harmonizing the created metadata and facilitating information integration.

Publication III gives a detailed view on how the query expansion facilities of the ontology service are used in practice in information retrieval scenarios. The query expansion widget uses the semantic relations of the ontology to refine the query with additional query terms to increase the recall of the search. For example, if the user is searching for ”animals”, the query can be expanded to include also ”cats” and ”dogs” based on the concept hierarchy.

As a proof of concept for the process of converting a legacy thesauri into the SKOS data model and publishing it in the ONKI SKOS service, the case of Finnish General Thesaurus YSA, is reported. YSA has been developed in the National Library of Finland since 1987 and is widely used in libraries, museums, and archives in Finland. In addition to YSA, over 80 national and international vocabularies, taxonomies, and ontologies are published in the ONKI SKOS system.

The user interface of the ONKI service, including the ontology directory, search view, and ontology browser have been developed in an iterative process where different versions of the system have been published and made accessible online as ONKI1 (Publication I), ONKI2 [258], ONKI3 [12], and ONKI Light [248]. The ONKI system has been run as a living lab ontology service since the official announcement in the autumn of 2008,

and before the successor service, Finto[1] of the National Library of Finland, was publicly released in January 2014, the ONKI service had over 10 000 monthly users (excluding the widget and API users), with 400 registered user domains for the widget and API use.

3.2 Publishing Multiple Ontologies (RQ2)

The research question 2 extends the research question 1 by introducing the dimension of multiple ontologies and ontology repositories used simultaneously.

2. How can a collection of independent or interconnected ontologies—in different formats and repositories—be published and utilized using shared user interfaces and APIs?

Such a multi-ontology scenario is prevalent in many cases, for example when indexing content with more than one ontology, or when trying to find and choose a relevant ontology for a specific use case. The publications IV and V present methods for solving issues in such use cases.

Publication IV discusses the publication of an ontology cloud consisting of individual, interconnected ontologies. The system is based on an upper ontology and domain-specific ontologies extending it. The publication process involves merging the component ontologies into a single, coherent representation, and using the ONKI service to publish it to end users. The structure of the ontology cloud appears as a single whole to the users, without emphasizing the ontology boundaries. Thus, the users should be able to use it in a straightforward way, for example in content indexing, focusing only on the main task of choosing relevant concepts, not ontologies.

The motivation for building such an ontology cloud is to facilitate the data integration of heterogeneous datasets from different domains, using domain-specific ontologies. The approach complements other existing techniques in data integration on the Semantic Web—data entity linking in Linked Open Data (LOD) and metadata schema linking in Linked Open Vocabularies (LOV). The formation of an ontology cloud can be more efficient since the mappings between ontologies can be reused for different datasets. Using an upper ontology as the base for the cloud instead of mapping individual ontologies on a one-to-one basis eliminates redundant

[1] `http://finto.fi`

mapping work.

For building such an ontology cloud for end users, ontology development and management processes need to take into account the semantic dependencies between the individual ontologies. This includes the identification of conceptually overlapping parts of the ontologies to avoid redundant development work, and the communication of the changes of the upper ontology to the domain ontologies extending it. The feasibility of the approach was demonstrated in practice by building the ontology cloud KOKO of more than 47,000 concepts, based on the Finnish General Finnish Ontology YSO and 15 domain ontologies. The ontologies were developed by a network of domain experts, where each organization took responsibility for managing a single domain. Based on the experiences in building KOKO, a set of seven principles guiding the building and management process of the cloud, was formed. The principles are general enough to be applied to other ontology clouds as well. The principles aim to streamline the ontology development process and ensure the semantic integrity of the resulting cloud, especially concerning the transitive subclass hierarchies of concepts. The developed approach of the proactive linking of ontologies as part of their development phase instead of mapping them afterwards aims to minimize redundant work and maximize interoperability.

Publication V discusses an environment consisting of several ontology repositories, where users need to access the repositories concurrently. The proposed Normalized Ontology Repository (NOR) approach allows accessing ontology repositories using shared tools and user interfaces based on a) harmonizing the representation of concepts in ontologies by using the SKOS data model, and b) providing a uniform API that encompasses the general ontology access needs, such as functionalities for getting the metadata of ontologies in the repository, searching for concepts, and querying their properties.

Based on the approach, it is possible to give the user an overview of the ontologies available in a set of ontology repositories. This helps the user to choose a suitable ontology repository or a specific ontology for her needs. The user can browse the ontologies using a uniform browser view, eliminating the need for learning to use ontology-specific user interfaces. The NOR API and concept representation is an extra layer on an ontology repository, meaning the functionalities of the repository are not restricted in any way. When the user has found a suitable ontology, she can move from the uniform NOR browser to the possible own, specialized user inter-

face provided by the underlying ontology repository. This mechanism is motivated because different ontologies might benefit from access mechanisms of different kinds, such as user interfaces and APIs. The system also allows the simultaneous usage of public ontology repositories and private repositories of organizations.

To demonstrate the feasibility of NOR, it has been applied in real-life use scenarios. The ONKI ontology repository itself uses such an approach for providing a uniform user interface for searching and browsing over 80 vocabularies and ontologies published in the ONKI SKOS backends. The backends encompass ontologies in different RDF-based formats, such as SKOS and RDFS, which are accessible via an HTTP API that provides a concept search functionality and normalizes the representation of the concepts. The ONKI frontend provides users with a possibility to perform a global search to all the ontologies at the same time. The system also includes a directory listing of all the ontologies available and offers a faceted search view to them, so the user can find such ontologies, for example, whose subject domain is ”business” and are published by ”FinnONTO Consortium”. The directory listing is built using a uniform metadata representation of the ontologies in the system. ONKI Widget uses the NOR approach in an even more heterogeneous environment, by providing users with access to not only ONKI SKOS backends, but also to the ONKI Geo ontology server [131], offering access to the geographical ontology of Finnish contemporary place names. The approach has also been tested by implementing a metasearch prototype for accessing the ontologies of ONKI and BioPortal repositories simultaneously, and even for accessing more general data repositories than ontology repositories, the CultureSampo portal [178] and SAHA metadata editor [165].

The relationship between the ontology cloud and NOR methodologies is a complementary one. The ontology cloud enables interoperability on the ontology level, whereas the NOR approach is focused on compatibility on the ontology service level. Building an ontology cloud requires mapping effort between the domain ontologies and the general upper ontology, and thus makes evident the relations between datasets described using different domain ontologies. On the other hand, in the NOR approach the ontologies are presented using a harmonized data model, but there is no need to map the ontologies to each other on the concept level. Thus, the system can be used for simultaneously accessing and processing even a set of mutually unrelated ontologies with shared tools.

3.3 Complex Knowledge Organization Systems (RQ3)

The research question 3 concerns the applicability of the Semantic Web technologies to managing richer KOSs as ontologies and publishing them using ontology services.

3. How can richer knowledge organization systems, such as biological nomenclatures and classifications, be managed as an ontology and published using an ontology service?

As a case study, the publications VI–VIII present the modeling and management of biological names and classifications. The proposed solution is an ontology model for taxonomic concepts and their scientific and vernacular names. Publication VI presents the Taxon Meta-Ontology (TaxMeOn) model which is aimed for the following data: 1) species checklists and mappings between them, 2) vernacular name collections, and 3) the changes of scientific names and classifications, and differing opinions of taxonomic concepts based on biological research results. The model makes a distinction between the taxonomic concept and its name. Thus, they can be managed separately, and the nomenclatural changes and re-classifications of a concept can be tracked and managed.

The model is flexible in a sense that it is designed to be suitable for data with different levels of details. The simplest use case is to express a static list of taxon names, but the model also supports more complex needs of representing the changes of names and taxonomic concepts. As the model is based on Linked Data, it offers possibilities to link divergent data serving divergent purposes and detailed information with more general information.

An advantage of the model is its practicality and applicability to real-life use cases. This is demonstrated by applying it into three use cases: 1) publishing biological species checklists in an ontology service (27 lists, over 80,000 names), 2) collaborative management of vernacular names (ca. 26,000 taxa), and 3) management of individual scientific name changes resulting from biological research results (9 genera).

Publication VII presents the use case of applying TaxMeOn into modeling and publishing species checklists of scientific names, and compares the ontology model with storing the checklists in a legacy relational database. The model allows mapping of taxa between different checklists (e.g., based on their congruency) and representing and managing the changes in taxo-

nomic concepts, their classifications, and names in individual checklists. The main advantages of the ontology model as opposed to a traditional database are the linkability to other datasets, extendability of the data model, (re)usability of the data via standard publication mechanisms, and possibility to edit the data with standard RDF tools. This means that the ontologies can be utilized in applications using general ontology services, without the need to implement domain-specific access mechanisms for biological name collections.

The species checklists were published in the ONKI ontology service, facilitating their reuse via user interfaces and APIs. The ONKI browser interface can be used for searching and browsing taxa, finding currently valid names, and tracing the temporal changes in the scientific names. The ONKI autocompletion widget provides a way for integrating an access mechanism to checklists into user applications, e.g., a content management system. Furthermore, HTTP and SOAP APIs are available for programmatic access and a SPARQL endpoint for querying the ontologies.

Publication VIII gives a more thorough presentation on how the TaxMeOn model can be applied into management of vernacular names. The model provides a solution for managing the approval process of common names, supporting the temporal tracking of their changes via statuses and their time stamps. The system is used by the Finnish Biology Society Vanamo[2] to manage the Finnish names of vascular plants in a collaborative way. In the typical workflow, a new common name is first proposed for a plant, after which it can be accepted to be the recommended name, and finally it can be made an alternative if another recommended name is introduced later.

We present the complete workflow for managing a vernacular name ontology from a collaborative development of the ontology to publishing it as Linked Open Data and in an ontology service which makes it accessible to the general public. The ontology is available in machine-processable RDF format, with explicit semantics, e.g., the hierarchical relations are set between the plant URIs, facilitating data integration and information retrieval in cases where data is combined from heterogeneous sources. The plant name ontology helps harmonizing the terminology, which in turn enhances communication between various users. Application developers can utilize the ontology by using the plant name URIs for unambiguous referencing to plant species.

[2] `http://www.vanamo.fi`

The applicability of the TaxMeOn model for its most complex use case, the management of individual scientific name changes based on biological research, has been demonstrated with a test dataset representing changes in the taxonomic classification of Afro-tropical beetle family Eucnemidae in Publication VI. The family has gone through numerous taxonomic treatments. For example, the position of the species *Pterotarsus historio* in the taxonomic classification has changed 22 times and at least eight taxonomic concepts are associated to the genus *Pterotarsus*. The TaxMeOn representation of the dataset encompasses the different conceptions of a taxon (e.g., *Pterotarsus*), the temporal order of the changes, and the references to scientific publications whose results justify these changes. Such a detailed information source provides a unified view of a complex taxon, which can be beneficial even to the researchers of biology, as the details of taxa have traditionally been scattered across the original publications, and piecing them together can be difficult and time-consuming. The detailed data can be further linked to other datasets with less taxonomic information, such as species checklists, which provides their users with more precise information.

3.4 Summary

To summarize the results, the research questions are re-visited in this section.

1. How can lightweight ontologies be published on the Semantic Web so that they can be utilized in content indexing and information retrieval tasks?

To faciliate the use of ontologies, they should be published as ontology services to fulfill the needs of different user groups, supporting their workflows during the phases of the life cycle of an ontology, e.g., in content indexing and information retrieval. For the cost-efficient reuse of the ontologies, user interface components and APIs can be used to integrate ontology-based functionalities into applications. The common functionalities for ontology use include concept search, browsing, and selection. W3C's SKOS data model can be used for harmonizing different legacy thesauri, which facilitates their publication via shared mechanisms in ontology services. Ontologies of different kinds might benefit from differing user interfaces, e.g., a thesaurus can be visualized as a tree, whereas a

geographical ontology might employ a map view.

2. How can a collection of independent or interconnected ontologies—in different formats and repositories—be published and utilized using shared user interfaces and APIs?

A set of interconnected ontologies can be combined into an ontology cloud that can be made accessible to end users via an ontology service as a single representation encompassing the different domains of the member ontologies. For building such an ontology cloud, ontology development and management processes need to take into account the semantic dependencies between the individual ontologies. The ontology development can be streamlined by formulating practical ontology design principles that guide the work and ensure the consistency of the cloud. The use of multiple ontology repositories simultaneously can be accomplished by using a shared upper data model for the ontologies, e.g., SKOS, and providing a shared API for accessing the ontologies. The approach facilitates, e.g., the building of aggregated ontology directories and global search on a network of ontology repositories.

3. How can richer knowledge organization systems, such as biological nomenclatures and classifications, be managed as an ontology and published using an ontology service?

Designing an ontology model for the specific needs of the domain, and mapping it to a harmonizing data model, e.g. SKOS, allows the publication of the ontology in shared ontology services. This way, the ontology can be accessed with general ontology user interfaces and APIs, without losing the detailed representation of the information. In the case study of biological names and classifications, the main modeling solutions in the TaxMeOn model are the separation of the taxonomic concepts and names, representation of the changes and their temporal order, and mappings between the different conceptions of the taxonomic concepts. The management and publication workflow of the ontology can be implemented using a general RDF editor and ontology service.

4. Discussion

Traditional, comparative evaluation of the models, processes, and tools developed in this thesis is difficult, as the systems provide novel solutions for the problems described in Chapter 1. In particular, evaluation of Semantic Web applications is difficult as the usefulness and usability of the systems depend on multiple factors: the quality of the heterogeneous source data used, the underlying search and inference software, and the user interface [265]. In more mature fields of computer science, such as in relational databases or text-based information retrieval, established data models and search algorithms are often available to be used as building blocks for creating new methods. In Semantic Web, such existing components are rare, and they usually have to be designed for every application. The state adds complexity to the development process and requires a level of maturity from the system to be properly evaluated. Unless all the components are of high quality, the system is not useful for the user.

As the evaluation method in this thesis, the extensive application to practice has been used as a proof of concept with the developed artifacts being adjusted based on real-life experiences in accordance with the principles of action research. Burstein and Gregor [42] have proposed criteria for evaluating systems development research, covering the following aspects: 1) theoretical and practical significance, 2) internal validity, 3) external validity, 4) objectivity, and 5) reliability. In the following, the research of this thesis is evaluated according to these criteria.

4.1 Theoretical Implications

4.1.1 Ontology Services (RQ1)

In comparison to the earlier ontology server research presented in Chapter 2, the main contribution of the ONKI ontology service model is the tight integration and support for the runtime use of ontologies, focusing on the content indexing and information retrieval use cases. Many previous efforts have been focusing on development and editing capabilities [20, 226, 43, 50, 231, 21, 213, 69, 49, 75, 92] of ontology servers, or merely publishing ontologies via ontology directories [235, 33, 113, 168, 122] or ontology browsers [40, 230] without providing modular components or web services that can be integrated into external applications. The developed system is domain-agnostic, general solution for publishing and using ontologies, and thus is not restricted to a single KOS or domain, as opposed to several other ontology servers. Compared with general semantic web search engines [67, 59, 198], the ONKI system provides focused support for ontology-based tasks, e.g., by displaying concept hierarchies. Ontology search engines [41, 217, 254], on the other hand, are tools suited especially for finding suitable ontologies, but not for using individual ontologies, e.g., in content indexing.

ONKI widget for integrating concept search and selection functionalities into external applications is similar to the widget models and implementations of the OCLC Terminology Services and BioPortal. The ONKI approach is more general than the OCLC widget, as ONKI widget can be integrated directly into the user interface of the application, and is not used as a separate browser toolbar. ONKI widget aims to provide a streamlined user experience, where the ontology-based functionalities are served with as little interference with the original user interface as possible. ONKI widget [268] was published earlier than the BioPortal widget [275], to the best of the author's knowledge[1]. The autocompletion component by Hildebrand et al. for general RDF repositories is based on similar ideas as ONKI widget. However, ONKI widget is focused on ontology-based interactions and is packaged as a ready-to-use service. One of the novelties of ONKI widget is the combination of the autocompletion

[1]The history of the NCBO Widgets (BioPortal) documentation page `https://www.bioontology.org/wiki/index.php/NCBO_Widgets` dates to 12 May 2009.

search and ontology browsing mechanism. Thus, the user can start the concept search task by typing in a query term, select a matching concept, and further refine the selection by browsing the ontology, e.g., to find more specific concepts based on the concept hierarchy. The system also supports semantic query expansion in similar vein as piloted in the FACET Web demonstrator and STAR project widgets, but has been made publicly available as a widget that can be integrated into external applications.

The HTTP and SOAP API of the ONKI service provide high level abstraction access to ontologies with a compact API specification, focusing on supporting concrete use cases of content indexing and information retrieval. The APIs are web-based, meaning they can be utilized in distributed systems, promoting loosely coupled services and complying with the resource- [78] and service-oriented [70] architectures. The use of the APIs is not tied to a single ontology modeling or programming language, as opposed to implementations such as SKOS API and OWL API, or general RDF libraries, such as Jena and RDFLib. Compared with the general RDF query language SPARQL, ONKI is focused on providing ontology-based functionalities on a more abstract level.

The common functionalities of an ontology server identified in this study based on the analysis of the user requirements of KOS—the concept search, browsing, and selection—are supported by previous literature [60, 108, 18, 95]. Also, the decision of using the SKOS vocabulary as the harmonizing model for publishing KOSs is reaffirmed by existing research [261, 237, 54, 50].

4.1.2 Publishing Multiple Ontologies (RQ2)

The ontology cloud model presented in this thesis is based on the idea of expressing multiple interlinked ontologies as a single coherent system, published in an ontology service. Previous research on the inter-ontology support in ontology servers has been focused on the creation and representation of concept mappings between individual ontologies [196, 172, 224, 264, 156, 190], and not providing them as a shared, easy-to-use, cross-domain ontology for use cases such as content indexing. The model is supported by previous research on upper ontologies [184, 170, 194] and ontology modularization [243, 7], where the ontology content is divided into subsets based on the generality and domain of the concepts.

The presented principles and methods for the creation and management of the ontology cloud complement the previous guidelines for ontology

interoperability [235] and general ontology design principles [99, 260, 101, 191, 273, 89], by emphasizing the importance of the consistency of the concept hierarchy. The proposed methods and tools for identifying the overlappings of the participating ontologies and propagating the changes of the upper ontology to the domain ontologies aim to ensure that the ontology cloud is valid, up-to-date, and easy to use. The cloud model is based on utilizing existing legacy thesauri, and as such the system maintains the backward compatibility with existing annotations, providing a cost-efficient way to publish legacy data as Linked Data.

The NOR approach of accessing multiple ontology repositories simultaneously is based on a distributed architecture, whereas many previous ontology server models are centralized services. The system is not tied to a single ontology server implementation, but is based on a shared API instead. The NOR API is comprised of access methods to ontologies and concepts on a high abstraction level, with a focus on the concept search and representation of concept information and ontology metadata. The API is not based on a specific ontology language, as opposed to lower level APIs, such as SKOS API and OWL API. On other hand, in contrast to general knowledge and agent communication languages, such as OKBC, FIPA-ACL, KQML, and S-APL, NOR API is focused on practical use cases of content indexing and information retrieval. To avoid the building of extraneous wrappers, which are used in many federated ontology access systems [8, 90], NOR is designed to be lightweight and simple in order to be easy to implement in ontology servers. The approach of using a highly abstracted API and harmonizing metadata model is similar to the more recent approaches of Ontohub and JSKOS-API, of which the latter uses the same SKOS data model and basic methods of entity search and lookup as NOR API. The ontology metadata used in NOR utilizes the existing metadata models VoID [13], Dublin Core, and FOAF [38], complements them by adding information of the NOR endpoint address, and can be extended with other ontology and dataset description vocabularies, such as Ontology Metadata Vocabulary (OMV) [109] and Data Catalog Vocabulary (DCAT) [174].

4.1.3 Complex Knowledge Organization Systems (RQ3)

The developed TaxMeOn model for managing biological nomenclatures and classifications is an example of how a rich KOS can be modeled and maintained as an ontology and published in an ontology service. Compared

with traditional species databases that aim to aggregate taxon names from various sources and harmonize their usage [223, 207, 225, 76, 94, 215, 64], TaxMeOn provides an explicit data model that is a general solution for managing heterogeneous biological name collections, and not tied to a single database system. As the model is based on Linked Data technologies, the data model can be extended in a flexible way and integrated with other data sources. The use of URIs as global identifiers in TaxMeOn is supported by the previous research which has emphasized the need for persistent identifiers when referring to taxa [25, 200, 209, 225, 205, 219], either in the form of a taxon name combined with a literature reference or a technical string.

Compared with previous ontology models on taxonomic information [228, 85, 86, 48], TaxMeOn is focused on practical name management of species checklists, research results, and other nomenclatural collections. The model supports the management of parallel classifications and nomenclatural conceptions of taxa in a semantically rich way, whereas some of the previous research have concentrated on modeling a single, unchangeable classification [228, 76]. Many Linked Data publishing projects of biological data and biodiversity data models [61, 251, 125, 220] are not focused on semantic rigor and therefore do not promote the machine processability of the contents optimally.

4.1.4 Summary

The theoretical implications of the methods and tools presented in this thesis are summarized in Table 4.1, by re-visiting the objectives of the ontology services defined in Section 1.2.

4.2 Practical Implications

4.2.1 Ontology Services (RQ1)

ONKI service provides out-of-the-box support for publishing and utilizing SKOS vocabularies and other lightweight ontologies in, e.g., content indexing, without needing to implement application specific user interfaces for end users. The system caters for many common, sharable tasks in ontology-based applications related to, e.g., concept finding, browsing, selecting, and query expansion. Lots of work and costs can be saved by implementing

Objective	Methodological and technological solutions
Ontology publication channel	General publication channel for various ontologies, created by different parties, including user-uploaded ontologies and ones fetched from external sources by automated processes.
Heterogeneous ontologies	Support for ontologies in different RDF-based formats, not tied to a single domain, modeling language, or a publisher.
Tools for metadata creation	Widgets and APIs that can be integrated into applications to support ontology-based cataloging practices.
Support for distributed content creation	Public living lab ontology service that can be used by a heterogeneous network of memory organisations for creating interoperable metadata based on shared ontologies.
Facilitate search tasks	The widgets and APIs support information retrieval by providing query expansion and cross-language search facilities.
Multiple ontologies and repositories	Publication of interlinked ontologies as a coherent, cross-domain cloud, and ability to perform federated search and uniform access to multiple ontology repositories based on a harmonized data model and shared API.
Programmatic access	HTTP and SOAP APIs for common ontology-based tasks, including concept search and lookup.
Evaluation by applying into practice	The feasibility of the developed ontology services has been demonstrated by extensive piloting in diverse use cases.
Promote complex KOSs	Rich knowledge organization systems can be mapped to a harmonizing data model, such as SKOS, and published in shared ontology services. The management of such a KOS can be realized using a general RDF editor.

Table 4.1. The objectives of the ontology services and the corresponding solutions presented in this thesis.

such functionalities in standard ways and providing them for production use as ready-to-use services. In this way, the use patterns of utilizing vocabularies in the user interface can be harmonized, which makes the systems easier to learn and use, as there is no need for vocabulary-specific access mechanisms. The ONKI service provides simple, yet powerful APIs and an autocompletion widget for content indexing and query expansion. The services can be used not only in ontology-based applications, but in legacy systems, which can utilize the ontologies in a similar way as traditional thesauri.

ONKI has been run as a living lab service since 2008 and has acted with the KOKO ontology cloud as the backbone of the Finnish national ontology infrastructure [136], which aims to enhance the interoperability of the collections of museums, libraries, archives, companies, and other organizations. Also, several international ontologies, such as Iconclass[2], Medical Subject Headings (MeSH)[3], the United Nations Standard Products and Services Code (UNSPSC)[4], and Integrated Public Sector Vocabulary

[2] `http://www.iconclass.nl`

[3] `http://www.nlm.nih.gov/mesh/`

[4] `http://www.unspsc.org`

(IPSV)[5], have been published in ONKI to facilitate their use in applications. The ONKI service has been used in the distributed ontology-based content creation approach of several semantic portals, such as CultureSampo, HealthFinland [247], and BookSampo [177]. The ontology infrastructure has gained maturity in Finland, and through technology transfer, ONKI's production level successor, Finto service has been run by the National Library of Finland since 2014 [249].

4.2.2 Publishing Multiple Ontologies (RQ2)

The design principles and tools presented in this thesis for building and managing a cloud of interlinked ontologies have been used to build the cross-domain KOKO cloud of Finnish ontologies. The ontologies are based on existing, established thesauri that have been used in various organizations for describing heterogeneous contents. The users of the legacy thesauri have been shifting to use the KOKO cloud, which facilitates the cross-domain interoperability of their datasets, as the novelties of KOKO include the mappings between the individual domain ontologies and the semantic consistency of the concept hierarchies. The maintenance and further development of the KOKO cloud have been transferred to the National Library of Finland, where it is managed by a network of domain ontology developers using the ontology cloud design principles and tools that are being further developed by the library.

The NOR approach of accessing multiple ontology repositories has been used as the internal architecture of the ONKI service, where multiple ontology backend servers are accessed and their information is aggregated by the frontend server. The system enables global search on the distributed ontology network, facilitating the comparison of different ontologies and using multiple ontologies simultaneously, e.g., in content indexing. By publishing ontologies using a shared API and metadata format, the users can access the ontologies using common tools and interfaces, making it easier for them to start using new ontologies and ontology repositories, and integrating them into external applications. The workflow of the users is streamlined as they do not have to access the ontology repositories separately and be familiar with the repository-specific user interfaces, modeling solutions, and other features. For the ontology publishers, NOR aims to increase the visibility of the ontologies as it is easier to incorporate

[5] `http://id.esd.org.uk/IPSV`

them into services that aggregate ontologies from different sources, such as ontology directories. The approach is applicable also to other kinds of data sources in addition to ontology repositories.

4.2.3 Complex Knowledge Organization Systems (RQ3)

The developed TaxMeOn model is a practical solution for managing various kinds of biological name collections. Its design principles include using terminology that is established in biology, focusing only on taxonomic information, and supporting data of various levels of granularity and of alternative views, in order to be simple, yet flexible to use. Accompanying the data model, the research presented in this thesis has contributed by providing a collaborative ontology maintenance and publication workflow utilizing existing tools, such as SAHA metadata editor and ONKI ontology service. By following the approach, it is straightforward and cost-efficient to develop and publish new ontologies for public use. The use of HTTP URIs as identifiers instead of LSIDs that are used in many previous taxonomic databases [148, 51, 55, 225, 215, 64] simplifies the publishing process and use of the taxonomic nomenclature. The system relies on standard resolving and locating mechanisms of the web infrastructure, without the need to implement a specialized LSID resolver.

As a proof of concept of the ontology model, several species checklists of the Finnish Museum of Natural History have been converted into TaxMeOn ontologies and published in the ONKI service. Different stakeholders, such as environmental authorities or biodiversity researchers, can use the system for cataloging, finding, and integrating information from heterogeneous sources, enabling the use of unambiguous taxon references. The ability to link scientific and vernacular names together is useful especially in the citizen science context and information retrieval by laymen as non-professionals might not be familiar with scientific nomenclature.

4.2.4 Summary

The practical implications of the methods and tools presented in this thesis are summarized in Table 4.2, by discussing how the ontology services and models have been applied in the case studies.

Feature	Application in a case study
Creation of interoperable metadata for memory organizations and semantic portals	
Shared ontologies	Semantic interoperability of heterogeneous data sources by creating ontology-based metadata.
Support for legacy data	By basing the ontologies on existing thesauri backwards compatibility to previously created contents is achieved.
Tools for content indexing	Support for common cataloging workflows in a cost-effective way, minimizing manual work.
Building the Finnish national ontology infrastructure	
National level ontology services	Free public services support the creation of standardized metadata in a shared way by governmental agencies, companies, and other organizations.
Formation of the KOKO cloud	A single interlinked representation of a set of domain ontologies makes the cross-domain data integration possible, acting as semantic glue. The redundant work of ontology developers can be eliminated as the overlapping parts of individual ontologies are identified.
Management process of the KOKO cloud	The domain ontologies are kept up-to-date regarding the changes of the general upper ontology, ensuring the validity and consistency of the ontology cloud.
Harmonized data model and API	Multiple ontologies, possibly originating from different repositories can be used simultaneously with shared tools in ontology-based workflows, e.g., in content indexing.
Management and publication of biological name collections	
Support for diverse name resources	The data model can be used for various kinds of name collections, facilitating their interoperability and linking.
Temporal management	The changes and different interpretations of the names and classifications can be tracked and controlled, supporting the scientific processes of taxonomy.
Extendability and external linking	The data model can be expanded to support new use cases and linked to external data sources providing additional information.
Tools for complete workflow	The processes for creating, managing, publishing, and using name collections are supported by general RDF editors, ontology services, and the web infrastructure.

Table 4.2. The features of the ontology services and models presented in this thesis and their application in the case studies.

4.3 Reliability and Validity

The reliability refers to the consistency of the research process and its stability over time and across researchers and methods. The research questions and objectives of this study are stated in Chapter 1, and the research questions are revisited when the results of the study are presented in Chapter 3. The developed models and prototype systems are presented explicitly in Chapter 3, and discussed in more details in the individual publications that are included in the thesis. The objectivity of the research is ensured by describing the research methods, the ontologies used in the study, and the organizations involved. The author of the thesis has no competing interests regarding the research presented and does not recognize personal biases that might affect the process.

The internal validity concerns the achievement of the stated objectives of the research and requirements of the developed systems, the alternative methods, and the limitations of the research. The models and systems developed meet the objectives presented in Chapter 1 as discussed in Chapter 3 and Sections 4.1 and 4.2, as evidenced by applying them in real-world use cases. The ONKI service has been used for publishing several ontologies, integrating them into external applications for creating interoperable metadata of distributed collections of museums and other organizations, and run as a living lab for supporting the users, including content indexers, information searchers, ontology developers, and application developers. The developed ontology models and principles have been used for the creation and management of the ontology cloud KOKO and several biological nomenclatures. The system demonstrates its applicability to the simultaneous usage of multiple ontologies and ontology repositories. The models and software implementations have been compared with relevant related work.

The usability of the user interface of the ONKI system was evaluated and improved by Rami Alatalo [12]. Based on the conducted user survey and interviews focusing on professional content indexers, and heuristic evaluation of the ONKI2 user interface, a new version of the user interface was built. The resulting user interface ONKI3 gained a mean usability score of 48/100 in the System Usability Scale (SUS) [39], based on a user survey. Considering the varying use cases and needs of the target users, the score can be viewed as decent.

In September 2011, a user inquiry was conducted for the users of the

ONKI service using an online questionnaire, consisting of Likert scale and open-ended questions. The inquiry received 107 responses from Finnish libraries, museums, archives, research institutes, and government agencies. The preliminary results of the inquiry were presented in a FinnONTO project board meeting [259]. Below, key findings regarding the functionalities of ONKI are summarized, based on the respondents who answered that they use ONKI daily or weekly.

- 58 % of the users regard the functionalities of ONKI as good or excellent for their needs, whereas 40 % of them think that the functionalities are poor or satisfactory.
- 57 % of the users who use the concept search continuously or occasionally regard the functionalities of ONKI as good or excellent for their needs, whereas 38 % of them think that the functionalities are poor or satisfactory.
- 45 % of the users who the use the ontology browsing continuously or occasionally regard the functionalities of ONKI as good or excellent for their needs, whereas 36 % of them think that the functionalities are poor or satisfactory.
- 55 % of the users think that ONKI is as good or better than the VESA Web Thesaurus Service[6], as opposed to 28 % who think that ONKI is worse than VESA.

When comparing ONKI with VESA, it should be noted that the results give insight on the usefulness of the two systems in relation to each other, not on the absolute quality of the systems. Based on the responses, the most used functionality of ONKI is the concept search. Overall, the ONKI service was regarded as important.

While the ONKI service can be considered an abstract concept, the concrete implementation introduces some limitations. For example, the KOSs that are to be published in the system need to be serialized in the RDF format. RDF is a non-proprietary format, and converting data into it is straightforward in many cases, but some complex data models might require careful and laborious design work. The proposed content

[6]A majority of the Finnish libraries, museums, and other memory organizations have used VESA previously to access established Finnish thesauri maintained by the National Library of Finland. VESA has been since replaced by the Finto service.

indexing methods are mainly manual, meaning they require expensive human efforts, though the methods streamline existing manual indexing processes and thus aim to save costs. In addition to manual indexing, the ontologies published in the ONKI system can be used via APIs, which can be used as components in automatic annotation systems.

The external validity refers to the generalizability of the research results and congruency with prior theory. The applicability of the developed methods to other settings has been ensured by demonstrating representative, diverse use cases in the proof-of-concept systems. The ONKI service has been used as a publishing platform for national and international thesauri and ontologies, covering different domains, ontology modeling languages, owners, and users. In addition to lightweight ontologies, the ONKI APIs and autocompletion widget have been used for accessing more complex geographical ontologies and biological classifications, with use cases in content indexing and query expansion. The tool that was developed for detecting overlappings between ontologies has been used not only in building the KOKO cloud, but also for generating the first version of a combined ontology of legal concepts, based on three vocabularies in the field of the Finnish legislation [87]. The NOR approach has been tested in three different use cases, including a demonstration involving an external ontology repository. The TaxMeOn model for biological nomenclatures has been applied into three distinct types of name collections, each with specific requirements.

The systems presented in the thesis have been designed by taking the previous research into account. The key functionalities of the ONKI service and NOR API are supported by the previous work on ontology servers, as the user requirements and common tasks in ontology-based workflows identified in this thesis are similar to findings by other researchers. The principles of building the ontology cloud complement existing, general ontology design patterns. The TaxMeOn model aims to be a simple, yet semantically rich solution, taking inspiration both from more theoretical modeling approaches and practical publishing efforts of the previous work.

4.4 Recommendations for Further Research

The research presented in this thesis paves way for future work on several areas. Deeper understanding about the roles and requirements of different ontology user groups and the suitability of the developed interfaces and tools for them would require more thorough user evaluation. The work

includes testing of the user interfaces of the ONKI service and the suitability of the APIs for various use cases. The use of ontology services in information retrieval needs to be further studied, as more information is required for the optimal selection of the relations used for expanding the queries, and for improving the user interactions in the ONKI widget. Such insight can be gained by conducting a formal evaluation covering alternative options.

Concerning the publication of ontologies, it would be beneficial to make the different versions of an ontology available for the users, not only the latest version as in the ONKI service. By doing so, users would be able to compare the different versions of ontologies, analyze their changes, change frequencies, etc. Access to historical versions of ontologies is needed especially in situations where the user has content that has been indexed with an older version of the ontology and wishes to make it compatible with the current version. Implementing such functionalities would require further research on ontology versioning and visualization methods and their application to concrete use cases.

Regarding the management of a cloud of interlinked ontologies, the processes and tooling for tracking and communicating the changes of the upper ontology to domain ontologies need to be refined. In addition, a more formal process for the development and overall coordination of the ontology cloud would further guide the ontology developers and streamline their work. Research on methods for validating the consistency of the ontology cloud is needed in order to ensure the high quality of the cloud, e.g., to avoid logical errors when reasoning over the class hierarchies. The work of developing the management processes of the ontology cloud is currently underway in the National Library of Finland.

The NOR API that is used for accessing multiple ontology repositories simultaneously could be extended to allow more functionalities. In order to keep the basic API as simple and cost-effective to implement as possible, the extensions could be defined as optional modules that are not required to be implemented in every ontology repository participating in the NOR network. The possibilities for the extensions include more fine-grained ways to restrict the concept search, support for mappings between ontologies, and ranking of the search results, e.g., based on the ontology or repository they are included in.

To facilitate the development and maintenance of biological nomenclatures using the TaxMeOn model, user-friendly tools are needed, as the

existing, generic RDF editors do not support efficient management of such complex ontologies. The user interface of the tool should hide the complexities of the data model and present the data in an intuitive and established way to biologists. To this end, research on developing a configurable RDF editor that can be adjusted to different data models would be beneficial. The TaxMeOn model could be extended with new structures to enable a more fine-grained modeling of hybrid taxa and taking into account the distinct features of zoological and botanical nomenclature. The value of the species checklists and name collections published in the ONKI service could be added by developing or using mapping tools to generate links from taxon names to complementing datasets, such as DBpedia.

Bibliography

[1] ISO 5963: Documentation – methods for examining documents, determining their subjects, and selecting indexing terms. ISO standard, International Organization for Standardization, 1985.

[2] SFS 5471: Guidelines for the establishment and maintenance of Finnish language thesauri. SFS standard, Finnish Standards Association, 1988.

[3] FIPA agent communication language specifications. Tech. rep., Foundation for Intelligent Physical Agents, 2001. `http://www.fipa.org/repository/aclspecs.html`.

[4] FIPA ontology service specification. Experimental specification, Foundation for Intelligent Physical Agents, 2001. `http://www.fipa.org/specs/fipa00086/XC00086D.html`.

[5] Global Strategy for Plant Conservation: Technical Rationale, Justification for Updating and Suggested Milestones and Indicators. Meeting report UNEP/CBD/COP/10/19, Convention on Biological Diversity, 2010.

[6] ISO 25964-1: Information and documentation – thesauri and interoperability with other vocabularies – part 1: Thesauri for information retrieval. ISO standard, International Organization for Standardization, 2011.

[7] ABBÈS, S. B., SCHEUERMANN, A., MEILENDER, T., AND D'AQUIN, M. Characterizing modular ontologies. In *Proceedings of the 6th International Workshop on Modular Ontologies, the 7th International Conference on Formal Ontology in I nformation Systems (FOIS 2012), Graz, Austria, July 24* (2012), T. Schneider and D. Walther, Eds., CEUR Workshop Proceedings.

[8] ADAMUSIAK, T., BURDETT, T., KURBATOVA, N., VAN DER VELDE, K. J., ABEYGUNAWARDENA, N., ANTONAKAKI, D., KAPUSHESKY, M., PARKINSON, H., AND SWERTZ, M. A. OntoCAT – simple ontology search and integration in Java, R and REST/JavaScript. *BMC Bioinformatics 12* (2011).

[9] AHMAD, M. N., AND COLOMB, R. M. Managing ontologies: A comparative study of ontology servers. In *Database Technologies 2007: Proceedings of the Eighteenth Australasian Database Conference (ADC2007), Ballarat, Victoria, Australia, January 29 - February 2* (2007), J. Bailey and A. Fekete, Eds., Australian Computer Society, pp. 13–22.

[10] AITCHISON, J., AND DEXTRE CLARKE, S. The Thesaurus: A Historical Viewpoint, with a Look to the Future. *Cataloging and Classification Quarterly 37*, 3/4 (2004), 5–21.

[11] AITCHISON, J., GILCHRIST, A., AND BAWDEN, D. *Thesaurus Construction and Use: A Practical Manual*, 4 ed. Europa Publications, London, 2000.

[12] ALATALO, R. Ontologiapalvelun käyttöliittymän jatkokehitys (Further development of an ontology service user interface), Dec. 2010. Bachelor's thesis. Aalto University, Espoo, Finland.

[13] ALEXANDER, K., CYGANIAK, R., HAUSENBLAS, M., AND ZHAO, J. Describing linked datasets with the VoID vocabulary. W3C interest group note, World Wide Web Consortium, 03 March 2011. http://www.w3.org/TR/2011/NOTE-void-20110303/.

[14] AMIN, A., HILDEBRAND, M., VAN OSSENBRUGGEN, J., EVERS, V., AND HARDMAN, L. List, group or menu: organizing suggestions in autocompletion interfaces. Tech. Rep. INS-E0901, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, Netherlands, 2009.

[15] ANDERSON, J. D. Guidelines for indexes and related information retrieval devices. Tech. rep., National Information Standards Organization (NISO), 1997.

[16] ANDREWS, P., ZAIHRAYEU, I., AND PANE, J. A Classification of Semantic Annotation Systems. *Semantic Web 3*, 3 (2012), 223–248.

[17] BACA, M., Ed. *Introduction to Metadata*, 3 ed. Getty Publications, Los Angeles, California, 2016. http://www.getty.edu/publications/intrometadata.

[18] BACLAWSKI, K., AND SCHNEIDER, T. The open ontology repository initiative: Requirements and research challenges. In *Proceedings of the Workshop on Collaborative Construction, Management and Linking of Structured Knowledge (CK2009), collocated with the 8th International Semantic Web Conference (ISWC-2009), Washington D.C., USA, October 25* (2009), T. Tudorache, G. Correndo, N. Noy, H. Alani, and M. Greaves, Eds., CEUR Workshop Proceedings.

[19] BAKER, T., BECHHOFER, S., ISAAC, A., MILES, A., SCHREIBER, G., AND SUMMERS, E. Key choices in the design of Simple Knowledge Organization System (SKOS). *Journal of Web Semantics 20* (2013), 35–49.

[20] BANDHOLTZ, T., SCHULTE-COERNE, T., GLASER, R., FOCK, J., AND KELLER, T. iQvoc – Open Source SKOS(XL) Maintenance and Publishing Tool. In *Proceedings of the Sixth Workshop on Scripting and Development for the Semantic Web, co-located with the European Semantic Web Conference 2010 (ESWC 2010), Crete, Greece, May 31* (2010), G. A. Grimnes, S. Auer, and G. T. Williams, Eds., CEUR Workshop Proceedings.

[21] BASCA, C., CORLOSQUET, S., CYGANIAK, R., FERNÁNDEZ, S., AND SCHANDL, T. Neologism: Easy vocabulary publishing. In *Proceedings of the 4th Workshop on Scripting for the Semantic Web, Tenerife, Spain, June 02* (2008), C. Bizer, S. Auer, G. A. Grimnes, and T. Heath, Eds., CEUR Workshop Proceedings.

[22] BASKAUF, S. J., AND WEBB, C. O. Darwin-SW: Darwin Core-based terms for expressing biodiversity data as RDF. *Semantic Web 7*, 6 (2016), 645–667.

[23] BASKERVILLE, R. L. Distinguishing action research from participative case studies. *Journal of Systems and Information Technology 1*, 1 (1997), 25–45.

[24] BASKERVILLE, R. L., AND WOOD-HARPER, A. T. A critical perspective on action research as a method for information systems research. *Journal of Information Technology 11*, 3 (1996), 235–246.

[25] BERENDSOHN, W. G. The concept of "potential taxa" in databases. *Taxon 44* (1995), 207–212.

[26] BERENDSOHN, W. G. A taxonomic information model for botanical databases: The IOPI model. *Taxon 46*, 2 (1997), 283–309.

[27] BERNERS-LEE, T. Linked data, W3C design issues. `http://www.w3.org/DesignIssues/LinkedData.html`, 2006.

[28] BERNERS-LEE, T., HENDLER, J., AND LASSILA, O. The semantic web. *Scientific American 284*, 5 (2001), 34–43.

[29] BHOGAL, J., MACFARLANE, A., AND SMITH, P. A review of ontology based query expansion. *Information Processing and Management 43*, 4 (2007), 866–886.

[30] BINDING, C., AND TUDHOPE, D. KOS at your Service: Programmatic Access to Knowledge Organisation Systems. *Journal of Digital Information 4*, 4 (2004).

[31] BINDING, C., AND TUDHOPE, D. Terminology Web Services. *Knowledge Organization 37*, 4 (2010), 287–298.

[32] BINDING, C., AND TUDHOPE, D. Improving interoperability using vocabulary linked data. *International Journal on Digital Libraries 17*, 1 (2016), 5–21.

[33] BLOMQVIST, E., GANGEMI, A., AND PRESUTTI, V. Experiments on pattern-based ontology design. In *Proceedings of the fifth international conference on Knowledge capture (K-CAP '09), Redondo Beach, California, USA, September 1-4* (2009), ACM, pp. 41–48.

[34] BODNER, R. C., AND SONG, F. Knowledge-Based Approaches to Query Expansion in Information Retrieval. In *Advances in Artifical Intelligence: 11th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, AI '96 Toronto, Ontario, Canada, May 21–24, 1996 Proceedings* (1996), G. McCalla, Ed., Springer-Verlag, pp. 146–158.

[35] BORST, W. N. *Construction of engineering ontologies for knowledge sharing and reuse*. PhD thesis, University of Twente, Netherlands, Sept. 1997.

[36] BOYLE, B., HOPKINS, N., LU, Z., RAYGOZA GARAY, J. A., MOZZHERIN, D., REES, T., MATASCI, N., NARRO, M. L., PIEL, W. H., MCKAY, S. J., LOWRY, S., FREELAND, C., PEET, R. K., AND ENQUIST, B. J. The taxonomic name resolution service: an online tool for automated standardization of plant names. *BMC Bioinformatics 14* (2013).

[37] BRAUN, S., SCHMIDT, A., WALTER, A., AND ZACHARIAS, V. The Ontology Maturing Approach for Collaborative and Work Integrated Ontology Development: Evaluation Results and Future Directions. In *Proceedings of the First International Workshop on Emergent Semantics and Ontology Evolution (ESOE 2007), co-located with ISWC 2007 + ASWC 2007, Busan, Korea, November 12* (2007), L. L. Chen, P. Cudré-Mauroux, P. Haase, A. Hotho, and E. Ong, Eds., CEUR Workshop Proceedings, pp. 5–18.

[38] BRICKLEY, D., AND MILLER, L. FOAF vocabulary specification 0.99. Namespace document, 14 January 2014. `http://xmlns.com/foaf/spec/20140114.html`.

[39] BROOKE, J. SUS - A quick and dirty usability scale. In *Usability evaluation in industry*, P. W. Jordan, B. Thomas, I. L. McClelland, and B. Weerdmeester, Eds. Taylor & Francis, 1996.

[40] BRUGMAN, H., MALAISÉ, V., AND GAZENDAM, L. A Web Based General Thesaurus Browser to Support Indexing of Television and Radio Programs. In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy, May 22-28* (2006).

[41] BUITELAAR, P., EIGNER, T., AND DECLERCK, T. OntoSelect: A Dynamic Ontology Library with Support for Ontology Selection. In *Proceedings of the 3rd International Semantic Web Conference (ISWC2004), Demo Papers, Hiroshima, Japan, November 7-11* (2004).

[42] BURSTEIN, F., AND GREGOR, S. The Systems Development or Engineering Approach to Research in Information Systems: An Action Research Perspective. In *Proceedings of 10th Australasian Conference on Information Systems, Wellington, New Zealand, December 1-3* (1999), B. Hope and P. Yoong, Eds., pp. 122–134.

[43] CARACCIOLO, C., STELLATO, A., RAJBAHNDARI, S., MORSHED, A., JOHANSSEN, G., JAQUES, Y., AND KEIZER, J. Thesaurus maintenance, alignment and publication as linked data: the AGROVOC use case. *International Journal of Metadata, Semantics and Ontologies 7*, 1 (2012), 65–75.

[44] CARDILLO, E., FOLINO, A., TRUNFIO, R., AND GUARASCI, R. Towards the reuse of standardized thesauri into ontologies. In *Proceedings of the 5th International Conference on Ontology and Semantic Web Patterns (WOP'14), Riva del Garda, Italy, October 19* (2014), V. de Boer, A. Gangemi, K. Janowicz, and A. Lawrynowicz, Eds., CEUR Workshop Proceedings, pp. 26–37.

[45] CARPINETO, C., AND ROMANO, G. A Survey of Automatic Query Expansion in Information Retrieval. *ACM Computing Surveys 44*, 1 (2012).

[46] CATHRO, W. Metadata: An Overview. In *Proceedings of the Standards Australia Seminar: Matching Discovery and Recovery, Sydney and Melbourne, Australia, August 14-15* (1997).

[47] CHAUDHRI, V. K., FARQUHAR, A., FIKES, R., KARP, P. D., AND RICE, J. P. Open knowledge base connectivity 2.0.3. Proposed specification, OKBC working group, April 9, 1998. `http://www.ai.sri.com/~okbc/spec/okbc2/okbc2.html`.

[48] CHAWUTHAI, R., TAKEDA, H., WUWONGSE, V., AND JINBO, U. Presenting and Preserving the Change in Taxonomic Knowledge for Linked Data. *Semantic Web 7*, 6 (2016), 589–616.

[49] CHUNG, H., CHOI, J.-M., YI, J.-H., HAN, J., AND LEE, E.-S. A Construction of the Adapted Ontology Server in EC. In *Intelligent Data Engineering and Automated Learning — IDEAL 2000. Data Mining, Financial Engineering, and Intelligent Agents: Second International Conference Shatin, N.T., Hong Kong, China, December 13–15, 2000 Proceedings* (2002), K. S. Leung, L.-W. Chan, and H. Meng, Eds., Springer-Verlag, pp. 355–360.

[50] CONWAY, M., KHOJOYAN, A., FANA, F., SCUBA, W., CASTINE, M., MOWERY, D., CHAPMAN, W., AND JUPP, S. Developing a web-based SKOS editor. *Journal of Biomedical Semantics 7*, 5 (2016).

[51] COSTELLO, M. J., BOUCHET, P., BOXSHALL, G., FAUCHALD, K., GORDON, D., HOEKSEMA, B. W., POORE, G. C. B., VAN SOEST, R. W. M., STÖHR, S., WALTER, T. C., VANHOORNE, B., DECOCK, W., AND APPELTANS, W. Global Coordination and Standardisation in Marine Biodiversity through the World Register of Marine Species (WoRMS) and Related Databases. *PLoS ONE 8*, 1 (2013).

[52] CÔTÉ, R. G., JONES, P., APWEILER, R., AND HERMJAKOB, H. The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics 7* (2006).

[53] COWIE, J., AND LEHNERT, W. Information extraction. *Communications of the ACM 39*, 1 (1996), 80–91.

[54] COX, S. J. D., YU, J., AND RANKINE, T. SISSVoc: A Linked Data API for access to SKOS vocabularies. *Semantic Web 7*, 1 (2016), 9–24.

[55] CROFT, J., CROSS, N., HINCHCLIFFE, S., LUGHADHA, E. N., STEVENS, P. F., WEST, J. G., AND WHITBREAD, G. Plant Names for the 21st Century: The International Plant Names Index, a Distributed Data Source of General Accessibility. *Taxon 48*, 2 (1999), 317–324.

[56] CRYER, P., HYAM, R., MILLER, C., NICOLSON, N., TUAMA, E. O., PAGE, R., REES, J., RICCARDI, G., RICHARDS, K., AND WHITE, R. Adoption of persistent identifiers for biodiversity informatics: Recommendations of the GBIF LSID GUID task group, 6. November 2009. Tech. rep., Global Biodiversity Information Facility (GBIF), Copenhagen, Denmark, 2010. version 1.1, last updated 21 Jan 2010.

[57] DAMERON, O., NOY, N. F., KNUBLAUCH, H., AND MUSEN, M. A. Accessing and manipulating ontologies using web services. In *Proceedings of the ISWC 2004 Workshop on Semantic Web Services: Preparing to Meet the World of Business Applications, Hiroshima, Japan, November 8* (2004), D. Martin, R. Lara, and T. Yamaguchi, Eds., CEUR Workshop Proceedings.

[58] D'AQUIN, M., AND LEWEN, H. Cupboard – a place to expose your ontologies to applications and the community. In *The Semantic Web: Research and Applications: 6th European Semantic Web Conference, ESWC 2009 Heraklion, Crete, Greece, May 31–June 4, 2009 Proceedings* (2009), L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou, and E. Simperl, Eds., Springer-Verlag, pp. 913–918.

[59] D'AQUIN, M., MOTTA, E., SABOU, M., ANGELETOU, S., GRIDINOC, L., LOPEZ, V., AND GUIDI, D. Toward a new generation of semantic web applications. *IEEE Intelligent Systems 23*, 3 (2008), 20–28.

[60] D'AQUIN, M., AND NOY, N. F. Where to publish and find ontologies? a survey of ontology libraries. *Journal of Web Semantics 11* (2012), 96–111.

[61] DARWIN CORE TASK GROUP. Darwin Core. TDWG current standard, Biodiversity Information Standards (TDWG), 2009. http://www.tdwg.org/standards/450.

[62] DAVISON, R. M., MARTINSONS, M. G., AND KOCK, N. Principles of Canonical Action Research. *Information Systems Journal 14* (2004), 65–86.

[63] DCMI USAGE BOARD. DCMI metadata terms. DCMI recommendation, Dublin Core Metadata Initiative, 14.6.2012. http://dublincore.org/documents/2012/06/14/dcmi-terms/.

[64] DE JONG, Y., VERBEEK, M., MICHELSEN, V., BJØRN, P. D. P., LOS, W., STEEMAN, F., BAILLY, N., BASIRE, C., CHYLARECKI, P., STLOUKAL, E., HAGEDORN, G., WETZEL, F. T., GLÖCKLER, F., KROUPA, A., KORB, G., HOFFMANN, A., HÄUSER, C., KOHLBECKER, A., MÜLLER, A., GÜNTSCH, A., STOEV, P., AND PENEV, L. Fauna Europaea - all European animal species on the web. *Biodiversity Data Journal 2* (2014).

[65] DENGLER, J., BERENDSOHN, W. G., BERGMEIER, E., CHYTRÝ, M., DANIHELKA, J., JANSEN, F., KUSBER, W.-H., LANDUCCI, F., MÜLLER, A., PANFILI, E., SCHAMINÉE, J., VENANZONI, R., AND VON RAAB-STRAUBE, E. The need for and the requirements of EuroSL, an electronic taxonomic reference list of all European plants. *Biodiversity & Ecology 4* (2012), 15–24.

[66] DEXTRE CLARKE, S. G., AND ZENG, M. L. From ISO 2788 to ISO 25964: The evolution of thesaurus standards towards interoperability and data modeling. *Information Standards Quarterly 24*, 1 (2012), 20–26.

[67] DING, L., FININ, T., JOSHI, A., PAN, R., COST, R. S., PENG, Y., REDDIVARI, P., DOSHI, V., AND SACHS, J. Swoogle: a search and metadata engine for the semantic web. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management (CIKM '04), Washington, D.C., USA, November 8-13* (2004), ACM, pp. 652–659.

[68] DING, Y., AND FENSEL, D. Ontology Library Systems: The key to successful Ontology Re-use. In *Proceedings of the 1st Semantic Web Working Symposium (SWWS'01), Stanford University, California, USA, July 30 - August 1* (2001), I. F. Cruz, S. Decker, J. Euzenat, and D. L. McGuinness, Eds., pp. 93–112.

[69] DOMINGUE, J. Tadzebao and WebOnto: Discussing, browsing, and editing ontologies on the web. In *Proceedings of the 11th Workshop on Knowledge Acquisition, Modeling and Management, Banff, Alberta, Canada, April 18-23* (1998).

[70] DRAHEIM, D. The Service-Oriented Metaphor Deciphered. *Journal of Computing Science and Engineering 4*, 4 (2010), 253–275.

[71] DUVAL, E. Metadata Standards: What, Who & Why. *Journal of Universal Computer Science 7*, 7 (2001), 591–601.

[72] EFTHIMIADIS, E. N. Interactive query expansion: A user-based evaluation in a relevance feedback environment. *Journal of the American Society for Information Science and Technology 51*, 11 (2000), 989–1003.

[73] EL JERROUDI, Z., AND ZIEGLER, J. iMERGE: interactive ontology merging. In *EKAW 2008: 16th International Conference on Knowledge Engineering and Knowledge Management: Knowledge Patterns, Acitrezza, Italy, September 29 - October 2, Poster and Demo Proceedings* (2008), pp. 52–56.

[74] EUZENAT, J., AND SHVAIKO, P. *Ontology Matching*. Springer-Verlag, Berlin Heidelberg, 2007.

[75] FARQUHAR, A., FIKES, R., AND RICE, J. The Ontolingua Server: a tool for collaborative ontology construction. *International Journal of Human-Computer Studies 46*, 6 (1997), 707–727.

[76] FEDERHEN, S. The NCBI Taxonomy database. *Nucleic Acids Research 40*, D1 (2012), 136–143.

[77] FEIGENBAUM, L., HERMAN, I., HONGSERMEIER, T., NEUMANN, E., AND STEPHENS, S. The semantic web in action. *Scientific American 297*, 6 (2007), 90–97.

[78] FIELDING, R. T. *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, University of California, Irvine, USA, 2000.

[79] FIKES, R., AND FARQUHAR, A. Distributed Repositories of Highly Expressive Reusable Ontologies. *IEEE Intelligent Systems 14*, 2 (1999), 73–79.

[80] FININ, T., WEBER, J., WIEDERHOLD, G., GENESERETH, M., FRITZSON, R., MCKAY, D., MCGUIRE, J., PELAVIN, R., SHAPIRO, S., AND BECK, C. Specification of the KQML agent-communication language – plus example agent policies and architectures. Draft specification, DARPA Knowledge Sharing Initiative External Interfaces Working Group, June 15, 1993. `http://www.csee.umbc.edu/csee/research/kqml/papers/kqmlspec.pdf`.

[81] FLOURIS, G., MANAKANATAS, D., KONDYLAKIS, H., PLEXOUSAKIS, D., AND ANTONIOU, G. Ontology change: classification and survey. *The Knowledge Engineering Review 23*, 2 (2008), 117–152.

[82] FLUIT, C., SABOU, M., AND VAN HARMELEN, F. Supporting User Tasks through Visualisation of Light-weight Ontologies. *Handbook on Ontologies in Information Systems* (2003), 1–20.

[83] FOX, E. A. Lexical Relations: Enhancing Effectiveness of Information Retrieval Systems. *ACM SIGIR Forum 15*, 3 (1980), 5–36.

[84] FRANZ, N. M., CHEN, M., KIANMAJD, P., YU, S., BOWERS, S., WEAKLEY, A. S., AND LUDÄSCHER, B. Names Are Not Good Enough: Reasoning over Taxonomic Change in the Andropogon Complex. *Semantic Web 7*, 6 (2016), 645–667.

[85] FRANZ, N. M., AND PEET, R. K. Towards a language for mapping relationships among taxonomic concepts. *Systematics and Biodiversity 7*, 1 (2009), 5–20.

[86] FRANZ, N. M., AND THAU, D. Biological taxonomy and ontology development: scope and limitations. *Biodiversity Informatics 7* (2010), 45–66.

[87] FROSTERUS, M., TUOMINEN, J., AND HYVÖNEN, E. Facilitating Re-use of Legal Data in Applications—Finnish Law as a Linked Open Data Service. In *Legal Knowledge and Information Systems: JURIX 2014: The Twenty-Seventh Annual Conference* (2014), R. Hoekstra, Ed., IOS Press, pp. 115–124.

[88] FU, G., JONES, C. B., AND ABDELMOTY, A. I. Ontology Based Spatial Query Expansion in Information Retrieval. In *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2005, Agia Napa, Cyprus, October 31 - November 4, 2005, Proceedings Part II* (2005), R. Meersman and Z. Tari, Eds., Springer-Verlag, pp. 1466–1482.

[89] GANGEMI, A., AND PRESUTTI, V. Ontology Design Patterns. In *Handbook on Ontologies*, S. Staab and R. Studer, Eds., 2 ed. Springer-Verlag, 2009.

[90] GARCÍA-CASTRO, A., GARCÍA-CASTRO, L., VILLAVECES, J. M., CALDERÓN, G., AND HEPP, M. OLS2OWL. A repository management facility. In *Proceedings of the 11th International Protégé Conference, Amsterdam, Netherlands, June 23-26* (2009).

[91] GARSHOL, L. M. Metadata? Thesauri? Taxonomies? Topic Maps! Making Sense of it all. *Journal of Information Science 30*, 4 (2004), 378–391.

[92] GENNARI, J. H., OLIVER, D. E., PRATT, W., RICE, J., AND MUSEN, M. A. A Web-Based Architecture for a Medical Vocabulary Server. In *Proceedings of the 19th Annual Symposium on Computer Applications in Medical Care, New Orleans, Lousiana, USA, October 28 - November 1* (1995), R. M. Gardner, Ed., American Medical Informatics Association, pp. 275–279.

[93] GIUNCHIGLIA, F., AND ZAIHRAYEU, I. Lightweight ontologies. In *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu, Eds. Springer-Verlag, 2009, pp. 1613–1619.

[94] GLOBAL BIODIVERSITY INFORMATION FACILITY (GBIF). ChecklistBank. `http://github.com/gbif/checklistbank`. Referenced 12 March 2017.

[95] GOLUB, K., TUDHOPE, D., ZENG, M. L., AND ŽUMER, M. Terminology Registries for Knowledge Organization Systems: Functionality, Use, and Attributes. *Journal of the Association for Information Science and Technology 65*, 9 (2014), 1901–1916.

[96] GREENBERG, J., LOSEE, R., PÉREZ AGÜERA, J. R., SCHERLE, R., WHITE, H., AND WILLIS, C. HIVE: Helping Interdisciplinary Vocabulary Engineering. *Bulletin of the American Society for Information Science & Technology 37*, 4 (2011), 23–26.

[97] GROSS, A., PRUSKI, C., AND RAHM, E. Evolution of biomedical ontologies and mappings: Overview of recent approaches. *Computational and Structural Biotechnology Journal 14* (2016), 333–340.

[98] GRUBER, T. R. A translation approach to portable ontology specification. *Knowledge Acquisition 5*, 2 (1993), 199–220.

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

[100] GUARINO, N., OBERLE, D., AND STAAB, S. What Is an Ontology? In *Handbook on ontologies*, S. Staab and R. Studer, Eds., 2 ed. Springer-Verlag, 2009, pp. 1–17.

[101] GUARINO, N., AND WELTY, C. Evaluating Ontological Decisions with Ontoclean. *Communications of the ACM 45*, 2 (2002), 61–65.

[102] GULLI, A., AND SIGNORINI, A. The Indexable Web is More than 11.5 Billion Pages. In *Proceedings of the WWW '05 Special interest tracks and posters of the 14th international conference on World Wide Web, Chiba, Japan, May 10-14* (2005), ACM, pp. 902–903.

[103] HALPIN, H., HAYES, P. J., MCCUSKER, J. P., MCGUINNESS, D. L., AND THOMPSON, H. S. When owl:sameAs isn't the Same: An Analysis of Identity in Linked Data. In *The Semantic Web – ISWC 2010: 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Revised Selected Papers, Part I* (2010), P. S. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, and B. Glimm, Eds., Springer-Verlag, pp. 305–320.

[104] HAMEED, A., PREECE, A., AND SLEEMAN, D. Ontology reconciliation. In *Handbook on ontologies*, S. Staab and R. Studer, Eds. Springer-Verlag, 2004, pp. 231–250.

[105] HARDISTY, A., ROBERTS, D., AND BIODIVERSITY INFORMATICS COMMUNITY. A decadal view of biodiversity informatics: challenges and priorities. *BMC Ecology 13*, 16 (2013).

[106] HARPER, C. A., AND TILLETT, B. B. Library of Congress Controlled Vocabularies and Their Application to the Semantic Web. *Cataloging & Classification Quarterly 43*, 3-4 (2007), 47–68.

[107] HARRIS, S., AND SEABORNE, A. SPARQL 1.1 query language. W3C recommendation, World Wide Web Consortium, 21 March 2013. `http://www.w3.org/TR/2013/REC-sparql11-query-20130321/`.

[108] HARTMANN, J., PALMA, R., AND GÓMEZ-PÉREZ, A. Ontology Repositories. In *Handbook on Ontologies*, S. Staab and R. Studer, Eds., 2 ed. Springer-Verlag, 2009, pp. 551–571.

[109] HARTMANN, J., PALMA, R., SURE, Y., SUÁREZ-FIGUEROA, M. C., HAASE, P., GÓMEZ-PÉREZ, A., AND STUDER, R. Ontology metadata vocabulary and applications. In *On the Move to Meaningful Internet Systems 2005: OTM 2005 Workshops: OTM Confederated Internationl Workshops and Posters, AWeSOMe, CAMS, GADA, MIOS+INTEROP, ORM, PhDS, SeBGIS, SWWS, and WOSE 2005, Agia Napa, Cyprus, October 31 - November 4, 2005. Proceedi* (2005), R. Meersman, Z. Tari, and P. Herrero, Eds., Springer-Verlag, pp. 906–915.

[110] HARTUNG, M., GROSS, A., AND RAHM, E. COnto-Diff: Generation of complex evolution mappings for life science ontologies. *Journal of Biomedical Informatics 46*, 1 (2013), 15–32.

[111] HARTUNG, M., TERWILLIGER, J., AND RAHM, E. Recent Advances in Schema and Ontology Evolution. *Evolution 141*, 3 (2011), 149–190.

[112] HEATH, T., AND BIZER, C. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool, Palo Alto, California, 2011.

[113] HEFLIN, J., AND HENDLER, J. Dynamic Ontologies on the Web. In *Proceedings of the 17th National Conference on Artificial Intelligence, Austin, Texas, USA, July 30 - August 3* (2000), H. A. Kautz and B. W. Porter, Eds., AAAI Press / MIT Press, pp. 443–449.

[114] HEFLIN, J. D. *Towards the Semantic Web: Knowledge Representation in a Dynamic, Distributed Environment*. PhD thesis, University of Maryland, USA, 2001.

[115] HERBERT, K. G., GEHANI, N. H., PIEL, W. H., WANG, J. T. L., AND WU, C. H. BIO-AJAX: an extensible framework for biological data cleaning. *ACM SIGMOD Record 33*, 2 (2004), 51–57.

[116] HEVNER, A. R., MARCH, S. T., PARK, J., AND RAM, S. Design Science in Information Systems Research. *MIS Quarterly 28*, 1 (2004), 75–105.

[117] HILDEBRAND, M., VAN OSSENBRUGGEN, J., AMIN, A., AROYO, L., WIELEMAKER, J., AND HARDMAN, L. The design space of a configurable autocompletion component. Tech. Rep. INS-E0708, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, Netherlands, 2007.

[118] HILDEBRAND, M., VAN OSSENBRUGGEN, J., HARDMAN, L., AND JACOBS, G. Supporting subject matter annotation using heterogeneous thesauri, a user study in Web data reuse. Tech. Rep. INS-E0902, Centrum Wiskunde & Informatica, 2009.

[119] HILL, L., BUCHEL, O., JANÉE, G., AND ZENG, M. L. Integration of Knowledge Organization Systems into Digital Library Architectures. In *13th ASIST SIGICR Workshop, Reconceptualizing Classification Research* (2002), pp. 46–52.

[120] HINCHLIFF, C. E., SMITH, S. A., ALLMAN, J. F., BURLEIGH, J. G., CHAUDHARY, R., COGHILL, L. M., CRANDALL, K. A., DENG, J., DREW, B. T., GAZIS, R., GUDE, K., HIBBETT, D. S., KATZ, L. A., LAUGHINGHOUSE, H. D., MCTAVISH, E. J., MIDFORD, P. E., OWEN, C. L., REE, R. H., REES, J. A., SOLTIS, D. E., WILLIAMS, T., AND CRANSTON, K. A. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences 112*, 41 (2015), 12764–12769.

[121] HITZLER, P., AND VAN HARMELEN, F. A Reasonable Semantic Web. *Semantic Web 1*, 1-2 (2010), 39–44.

[122] HLAVA, M. Developing an eclectic terminology registry. *Bulletin of the Association for Information Science and Technology 37*, 4 (2011), 19–22.

[123] HOANG, H. H., AND TJOA, A. M. The State of the Art of Ontology-based Query Systems: A Comparison of Existing Approaches. In *Proceedings of the International Conference on Computing and Informatics, Kuala Lumpur, Malaysia, June 6-8* (2006), IEEE.

[124] HODGE, G. Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files. Tech. rep., The Digital Library Federation, Council on Library and Information Resources, 2000.

[125] HOLETSCHEK, J., DRÖGE, G., GÜNTSCH, A., AND BERENDSOHN, W. G. The ABCD of primary biodiversity data access. *Plant Biosystems - An International Journal Dealing with all Aspects of Plant Biology 146*, 4 (2012), 771–779.

[126] HOLLINK, L., SCHREIBER, G., AND WIELINGA, B. Patterns of semantic relations to improve image content search. *Journal of Web Semantics 5*, 3 (2007), 195–203.

[127] HOOI, Y. K., HASSAN, M. F., AND SHARIFF, A. M. A Survey on Ontology Mapping Techniques. In *Advances in Computer Science and its Applications: CSA 2013* (2014), H. Y. Jeong, M. S. Obaidat, N. Y. Yen, and J. J. J. H. Park, Eds., Springer-Verlag, pp. 829–836.

[128] HUANG, Z., AND STUCKENSCHMIDT, H. Reasoning with multi-version ontologies: A temporal logic approach. In *The Semantic Web – ISWC 2005: 4th International Semantic Web Conference, ISWC 2005, Galway, Ireland, November 6-10, 2005, Proceedings* (2005), Y. Gil, E. Motta, V. R. Benjamins, and M. A. Musen, Eds., Springer-Verlag, pp. 398–412.

[129] HYVÖNEN, E., ALONEN, M., KOHO, M., AND TUOMINEN, J. BirdWatch — Supporting Citizen Scientists for Better Linked Data Quality for Biodiversity Management. In *Proceedings of the first international Workshop on Semantics for Biodiversity Montpellier, France, May 27* (2013), P. Larmande, E. Arnaud, I. Mougenot, C. Jonquet, T. Libourel, and M. Ruiz, Eds., CEUR Workshop Proceedings.

[130] HYVÖNEN, E., IKKALA, E., AND TUOMINEN, J. Linked data brokering service for historical places and maps. In *Proceedings of the 1st Workshop on Humanities in the Semantic Web co-located with 13th ESWC Conference 2016 (ESWC 2016) Anissaras, Greece, May 29* (2016), A. Adamou, E. Daga, and L. Isaksen, Eds., CEUR Workshop Proceedings, pp. 39–52.

[131] HYVÖNEN, E., LINDROOS, R., KAUPPINEN, T., AND HENRIKSSON, R. An Ontology Service for Geographical Content. In *Poster Proceedings of the International Semantic Web Conference (ISWC 2007), Busan, Korea, November 11-15* (2007).

[132] HYVÖNEN, E., AND MÄKELÄ, E. Semantic Autocompletion. In *The Semantic Web – ASWC 2006: First Asian Semantic Web Conference, Beijing, China, September 3-7, 2006, Proceedings* (2006), R. Mizoguchi, Z. Shi, and F. Giunchiglia, Eds., Springer-Verlag, pp. 739–751.

[133] HYVÖNEN, E., MÄKELÄ, E., KAUPPINEN, T., ALM, O., KURKI, J., RUOTSALO, T., SEPPÄLÄ, K., TAKALA, J., PUPUTTI, K., KUITTINEN, H., VILJANEN, K., TUOMINEN, J., PALONEN, T., FROSTERUS, M., SINKKILÄ, R., PAAKKARINEN, P., LAITIO, J., AND NYBERG, K. CultureSampo – A National Publication System of Cultural Heritage on the Semantic Web 2.0. In *The Semantic Web: Research and Applications: 6th European Semantic Web Conference, ESWC 2009 Heraklion, Crete, Greece, May 31–June 4, 2009 Proceedings* (2009), L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano,

T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou, and E. Simperl, Eds., Springer-Verlag, pp. 851–856.

[134] HYVÖNEN, E., TUOMINEN, J., ALONEN, M., AND MÄKELÄ, E. Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets. In *The Semantic Web: ESWC 2014 Satellite Events: ESWC 2014 Satellite Events, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers* (2014), V. Presutti, E. Blomqvist, R. Troncy, H. Sack, I. Papadakis, and A. Tordai, Eds., Springer-Verlag, pp. 226–230.

[135] HYVÖNEN, E., TUOMINEN, J., KAUPPINEN, T., AND VÄÄTÄINEN, J. Representing and utilizing changing historical places as an ontology time series. In *Geospatial Semantics and the Semantic Web: Foundations, Algorithms, and Applications*, N. Ashish and A. P. Sheth, Eds. Springer-Verlag, 2011, pp. 1–25.

[136] HYVÖNEN, E., VILJANEN, K., TUOMINEN, J., AND SEPPÄLÄ, K. Building a National Semantic Web Ontology and Ontology Service Infrastructure—The FinnONTO Approach. In *The Semantic Web: Research and Applications: 5th European Semantic Web Conference, ESWC 2008, Tenerife, Canary Islands, Spain, June 1-5, 2008 Proceedings* (2008), S. Bechhofer, M. Hauswirth, J. Hoffmann, and M. Koubarakis, Eds., Springer-Verlag, pp. 95–109.

[137] HÖLSCHER, C., AND STRUBE, G. Web search behaviour of internet experts and newbies. *Computer networks 33*, 1–6 (2000), 337–346.

[138] ICZN - INTERNATIONAL COMMISSION ON ZOOLOGICAL NOMENCLATURE. *International Code of Zoological Nomenclature*. International Trust for Zoological Nomenclature, 1999.

[139] ISO TC46/SC9/WG8, AND ISAAC, A. Correspondence between ISO 25964 and SKOS/SKOS-XL Models. Tech. rep., National Information Standards Organization (NISO), 2013.

[140] JAIN, P., HITZLER, P., KUNAL VERMA, YEH, P. Z., AND SHETH, A. Moving beyond sameAs with PLATO : Partonomy detection for Linked Data. In *Proceedings of 23rd ACM conference on Hypertext and social media, Milwaukee, USA, June 25-28* (2012), ACM, pp. 33–42.

[141] JARRAR, M., AND MEERSMAN, R. Formal Ontology Engineering in the DOGMA Approach. In *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE: Confederated International Conferences CoopIS, DOA, and ODBASE 2002 Proceedings* (2002), R. Meersman and Z. Tari, Eds., Springer-Verlag, pp. 1238–1254.

[142] JÄRVELIN, K., KEKÄLÄINEN, J., AND NIEMI, T. ExpansionTool: Concept-Based Query Expansion and Construction. *Information Retrieval 4*, 3-4 (2001), 231–255.

[143] JASPER, R., AND USCHOLD, M. A Framework for Understanding and Classifying Ontology Applications. In *Proceedings of the Twelfth Workshop on Knowledge Acquisition, Modeling and Management, Banff, Alberta, Canada, October 16-22* (1999).

[144] JAVED, M., ABGAZ, Y. M., AND PAHL, C. Ontology Change Management and Identification of Change Patterns. *Journal on Data Semantics 2*, 2 (2013), 119–143.

[145] JETZ, W., MCPHERSON, J. M., AND GURALNICK, R. P. Integrating biodiversity distribution knowledge: Toward a global map of life. *Trends in Ecology and Evolution 27*, 3 (2012), 151–159.

[146] JIMÉNEZ RUIZ, E., CUENCA GRAU, B., HORROCKS, I., AND BERLANGA, R. Supporting concurrent ontology development: Framework, algorithms and tool. *Data & Knowledge Engineering 70*, 1 (2011), 146–164.

[147] JOHO, H., COVERSON, C., SANDERSON, M., AND BEAULIEU, M. Hierarchical presentation of expansion terms. In *Proceedings of the 2002 ACM symposium on Applied computing, Madrid, Spain, March 11-14* (2002), ACM, pp. 645–649.

[148] JONES, A. C., WHITE, R. J., AND ORME, E. R. Identifying and relating biological concepts in the Catalogue of Life. *Journal of Biomedical Semantics 2*, 7 (2011).

[149] JOTHILAKSHMI, R., SHANTHI, N., AND BABISARASWATHI, R. A survey on semantic query expansion. *Journal of Theoretical and Applied Information Technology 57*, 1 (2013), 128–138.

[150] JUDD, W. S., CAMPBELL, C. S., KELLOGG, E. A., STEVENS, P. F., AND DONOGHUE, M. J. *Plant Systematics: A Phylogenetic Approach*, 4 ed. Sinauer Associates, Sunderland, Massachusetts, 2016.

[151] JUPP, S., BECHHOFER, S., AND STEVENS, R. A flexible API and editor for SKOS. In *The Semantic Web: Research and Applications: 6th European Semantic Web Conference, ESWC 2009 Heraklion, Crete, Greece, May 31–June 4, 2009 Proceedings* (2009), L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou, and E. Simperl, Eds., Springer-Verlag, pp. 506–520.

[152] KATASONOV, A., AND TERZIYAN, V. Semantic Agent Programming Language (S-APL): A Middleware Platform for the Semantic Web. In *Proceedings of the IEEE International Conference on Semantic Computing 2008 (ICSC 2008), Santa Clara, California, USA, August 4-7* (2008), IEEE, pp. 504–511.

[153] KAUPPINEN, T., KUITTINEN, H., TUOMINEN, J., SEPPÄLÄ, K., AND HYVÖNEN, E. Extending an Ontology by Analyzing Annotation Cooccurrences in a Semantic Cultural Heritage Portal. In *Proceedings of the ASWC 2008 Workshop on Collective Intelligence (ASWC-CI 2008) organized as a part of the 3rd Asian Semantic Web Conference (ASWC 2008), Bangkok, Thailand, February 2-5* (2009), pp. 1–6.

[154] KEEN, M., ACHARYA, A., BISHOP, S., HOPKINS, A., MILINSKI, S., NOTT, C., ROBINSON, R., ADAMS, J., AND VERSCHUEREN, P. *Patterns: Implementing an SOA using an enterprise service bus*. IBM, 2004.

[155] KENNEDY, J. B., KUKLA, R., AND PATERSON, T. Scientific names are ambiguous as identifiers for biological taxa: their context and definition are required for accurate data integration. In *Data Integration in the Life*

Sciences: Second International Workshop, DILS 2005, San Diego, CA, USA, July 20-22, 2005, Proceedings (2005), B. Ludäscher and L. Raschid, Eds., Springer-Verlag, pp. 80–95.

[156] KHAN, Z. C., AND KEET, C. M. The foundational ontology library ROMULUS. In *Model and Data Engineering: Third International Conference, MEDI 2013, Amantea, Italy, September 25-27, 2013, Proceedings* (2013), A. Cuzzocrea and S. Maabout, Eds., Springer-Verlag, pp. 200–211.

[157] KHATTAK, A. M., LATIF, K., AND LEE, S. Change management in evolving web ontologies. *Knowledge-Based Systems 37* (2013), 1–18.

[158] KIRSTEN, T., GROSS, A., HARTUNG, M., AND RAHM, E. GOMMA: a component-based infrastructure for managing and analyzing life science ontologies and their evolution. *Journal of Biomedical Semantics 2*, 6 (2011).

[159] KIRYAKOV, A., POPOV, B., TERZIEV, I., MANOV, D., AND OGNYANOFF, D. Semantic annotation, indexing, and retrieval. *Journal of Web Semantics 2*, 1 (2004), 49–79.

[160] KLEIN, M. *Change Management for Distributed Ontologies*. PhD thesis, Vrije Universiteit Amsterdam, Netherlands, Aug. 2004.

[161] KLEIN, M., AND FENSEL, D. Ontology versioning on the Semantic Web. In *Proceedings of the 1st Semantic Web Working Symposium (SWWS'01), Stanford University, California, USA, July 30 - August 1* (2001), I. F. Cruz, S. Decker, J. Euzenat, and D. L. McGuinness, Eds., pp. 75–91.

[162] KLESS, D., JANSEN, L., AND MILTON, S. A content-focused method for re-engineering thesauri into semantically adequate ontologies using OWL. *Semantic Web 7*, 5 (2016), 543–576.

[163] KNAPP, P., POLASZEK, A., AND WATSON, M. Spreading the word. *Nature 446* (2007), 261–262.

[164] KOZAKI, K., SUNAGAWA, E., KITAMURA, Y., AND MIZOGUCHI, R. A framework for cooperative ontology construction based on dependency management of modules. In *Proceedings of the First International Workshop on Emergent Semantics and Ontology Evolution (ESOE 2007), co-located with ISWC 2007 + ASWC 2007, Busan, Korea, November 12* (2007), L. L. Chen, P. Cudré-Mauroux, P. Haase, A. Hotho, and E. Ong, Eds., CEUR Workshop Proceedings, pp. 33–44.

[165] KURKI, J., AND HYVÖNEN, E. Collaborative metadata editor integrated with ontology services and faceted portals. In *Proceedings of the 1st Workshop on Ontology Repositories and Editors for the Semantic Web, Hersonissos, Crete, Greece, May 31* (2010), M. d'Aquin, A. García Castro, C. Lange, and K. Viljanen, Eds., CEUR Workshop Proceedings, pp. 7–11.

[166] LAPP, H., MORRIS, R. A., CATAPANO, T., HOBERN, D., AND MORRISON, N. Organizing our knowledge of biodiversity. *Bulletin of the American Society for Information Science and Technology 37*, 4 (2011), 38–42.

[167] LEADBETTER, A. The NERC Vocabulary Server version 2.0. Tech. rep., British Oceanographic Data Centre, 2012.

[168] LEDL, A., AND VOSS, J. Describing Knowledge Organization Systems in BARTOC and JSKOS. In *Proceedings of the 12th International Conference on Terminology and Knowledge Engineering, Copenhagen, Denmark, June 22-24* (2016), H. Erdman Thomsen, A. Pareja-Lora, and B. Nistrup Madsen, Eds., pp. 168–178.

[169] LEENHEER, P. D., AND MENS, T. ONTOLOGY EVOLUTION: State of the Art & Future Directions. *Ontology Theory Management and Design-Advanced Tools and Models 2*, 1 (2010), 1–47.

[170] LENAT, D. B. CYC: A Large-Scale Investment in Knowledge Infrastructure. *Communications of the ACM 38*, 11 (1995), 33–38.

[171] LEPAGE, D., VAIDYA, G., AND GURALNICK, R. Avibase - A database system for managing and organizing taxonomic concepts. *ZooKeys 135*, 420 (2014), 117–135.

[172] LI, Y., THOMPSON, S., TAN, Z., GILES, N., AND GHARIB, H. Beyond ontology construction; ontology services as online knowledge sharing communities. In *Semantic Web - ISWC 2003: Second International Semantic Web Conference, Sanibel Island, FL, USA, October 20-23, 2003, Proceedings* (2003), D. Fensel, K. Sycara, and J. Mylopoulos, Eds., Springer-Verlag, pp. 469–483.

[173] LUGHADHA, E. N. Towards a working list of all known plant species. *Philosophical transactions of the Royal Society of London B: Biological sciences 359*, 1444 (2004), 681–687.

[174] MAALI, F., AND ERICKSON, J. Data catalog vocabulary (DCAT). W3C recommendation, World Wide Web Consortium, 16 January 2014. `http://www.w3.org/TR/2014/REC-vocab-dcat-20140116/`.

[175] MAEDCHE, A., MOTIK, B., AND STOJANOVIC, L. Managing multiple and distributed ontologies on the Semantic Web. *The International Journal on Very Large Data Bases (The VLDB Journal) 12*, 4 (2003), 286–302.

[176] MAEDCHE, A., AND STAAB, S. Ontology learning in the semantic web. *IEEE Intelligent Systems 16*, 2 (2001), 72–79.

[177] MÄKELÄ, E., HYPÉN, K., AND HYVÖNEN, E. BookSampo—Lessons Learned in Creating a Semantic Portal for Fiction Literature. In *The Semantic Web – ISWC 2011: 10th International Semantic Web Conference, Bonn, Germany, October 23-27, 2011, Proceedings, Part II* (2011), L. Aroyo, C. Welty, H. Alani, J. Taylor, A. Bernstein, L. Kagal, N. Noy, and E. Blomqvist, Eds., Springer-Verlag, pp. 173–188.

[178] MÄKELÄ, E., RUOTSALO, T., AND HYVÖNEN, E. How to deal with massively heterogeneous cultural heritage data – lessons learned in CultureSampo. *Semantic Web 3*, 1 (2012), 85–109.

[179] MALAISÉ, V., AROYO, L., BRUGMAN, H., GAZENDAM, L., DE JONG, A., NEGRU, C., AND SCHREIBER, G. Evaluating a thesaurus browser for an audio-visual archive. In *Managing Knowledge in a World of Networks: 15th International Conference, EKAW 2006, Poděbrady, Czech Republic, October 2-6, 2006, Proceedings* (2006), S. Staab and V. Svátek, Eds., Springer-Verlag, pp. 272–286.

[180] MANGOLD, C. A survey and classification of semantic search approaches. *International Journal of Metadata, Semantics and Ontologies 2*, 1 (2007), 23–34.

[181] MANNING, C. D., RAGHAVAN, P., AND SCHÜTZE, H. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[182] MARCH, S. T., AND SMITH, G. F. Design and natural science research on information technology. *Decision Support Systems 15*, 4 (1995), 251–266.

[183] MARTÍNEZ-GONZÁLEZ, M. M., AND ALVITE-DÍEZ, M.-L. On the evaluation of thesaurus tools compatible with the Semantic Web. *Journal of Information Science 40*, 6 (2014), 711–722.

[184] MASCARDI, V., CORDÌ, V., AND ROSSO, P. A Comparison of Upper Ontologies. In *Proceedings of the 8th Workshop Dagli Oggetti agli Agenti: Agenti e Industria: Applicazioni tecnologiche degli agenti software (WOA 2007), Genova, Italy, September 24-25* (2007), pp. 55–64.

[185] MAYDEN, R. L. A hierarchy of species concepts: the denouement in the saga of the species problem. In *Species: The Units of Biodiversity*, M. F. Oaridge, H. A. Dawah, and M. R. Wilson., Eds. Chapman & Hall, 1997, pp. 381–424.

[186] MAYNARD, D., PETERS, W., D'AQUIN, M., AND SABOU, M. Change management for metadata evolution. In *Proceedings of the International Workshop on Ontology Dynamics (IWOD-07), the 4th European Semantic Web Conference (ESWC-07), Innsbruck, Austria, June 7* (2007), G. Flouris and M. d'Aquin, Eds., pp. 27–40.

[187] MCNEILL, J., BARRIE, F. R., BUCK, W. R., DEMOULIN, V., GREUTER, W., HAWKSWORTH, D. L., HERENDEEN, P. S., KNAPP, S., MARHOLD, K., PRADO, J., PRUD'HOMME VAN REINE, W. F., SMITH, G. F., WIERSEMA, J. H., AND TURLAND, N. J. *International Code of Nomenclature for algae, fungi, and plants (Melbourne Code) adopted by the Eighteenth International Botanical Congress Melbourne, Australia, July 2011*. Koeltz Scientific Books, Oberreifenberg, 2012.

[188] MILES, A., AND BECHHOFER, S. SKOS simple knowledge organization system reference. W3C recommendation, World Wide Web Consortium, 18 August 2009. `http://www.w3.org/TR/2009/REC-skos-reference-20090818/`.

[189] MILES, A., MATTHEWS, B., WILSON, M., AND BRICKLEY, D. SKOS Core: Simple Knowledge Organisation for the Web. In *Proceedings of the International Conference on Dublin Core and Metadata Applications 2008, Madrid, Spain, September 12-15* (2005), Dublin Core Metadata Initiative, pp. 3–10.

[190] MOSSAKOWSKI, T., KUTZ, O., AND CODESCU, M. Ontohub: A semantic repository for heterogeneous ontologies. In *Proceedings of the Theory Day in Computer Science (DACS 2014), satellite workshop of ICTAC 2014, Bucharest, Romania, September 17-19* (2014).

[191] NAGYPÁL, G., AND MÜLLER, W. Ontology Design. In *Semantics At Work, Ontology Management – Tools and Techniques*, R. Volz, Ed. Lulu Press, 2008.

[192] NAVIGLI, R., AND VELARDI, P. An Analysis of Ontology-based Query Expansion Strategies. *Information Retrieval* (2002), 42–49.

[193] NEUBERT, J. Bringing the "thesaurus for economics" on to the web of linked data. In *Proceedings of the Linked Data on the Web Workshop (LDOW2009), Madrid, Spain, April 20* (2009), C. Bizer, T. Heath, T. Berners-Lee, and K. Idehen, Eds., CEUR Workshop Proceedings.

[194] NILES, I., AND PEASE, A. Towards a Standard Upper Ontology. In *Proceedings of the international conference on Formal Ontology in Information Systems (FOIS-2001), Ogunquit, Maine, USA, October 17 - 19* (2001), ACM, pp. 2–9.

[195] NOY, N. F., AND KLEIN, M. Ontology evolution: Not the same as schema evolution. *Knowledge and Information Systems 6*, 4 (2004), 428–440.

[196] NOY, N. F., SHAH, N. H., WHETZEL, P. L., DAI, B., DORF, M., GRIFFITH, N., JONQUET, C., RUBIN, D. L., STOREY, M.-A., CHUTE, C. G., AND MUSEN, M. A. BioPortal: Ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research 37*, Web Server issue (2009), W170–W173.

[197] OBERLE, D., STAAB, S., STUDER, R., AND VOLZ, R. Supporting application development in the semantic web. *ACM Transactions on Internet Technology 5*, 2 (2005), 328–358.

[198] OREN, E., DELBRU, R., CATASTA, M., CYGANIAK, R., STENZHORN, H., AND TUMMARELLO, G. Sindice.com: A Document-oriented Lookup Index for Open Linked Data. *International Journal of Metadata, Semantics and Ontologies 3*, 1 (2008), 37–52.

[199] OREN, E., MÖLLER, K. H., SCERRI, S., HANDSCHUH, S., AND SINTEK, M. What are Semantic Annotations? Tech. rep., DERI Galway, 2006.

[200] PAGE, R. D. M. Taxonomic names, metadata, and the Semantic Web. *Biodiversity Informatics 3* (2006), 1–15.

[201] PAGE, R. D. M. Biodiversity informatics: The challenge of linking data and the role of shared identifiers. *Briefings in Bioinformatics 9*, 5 (2008), 345–354.

[202] PAGE, R. D. M. Linking NCBI to Wikipedia: A wiki-based approach. *PLoS Currents* (2011), 1–14.

[203] PAGE, R. D. M. BioNames: linking taxonomy, texts, and trees. *PeerJ 1* (2013).

[204] PAN, J., CRANEFIELD, S., AND CARTER, D. A lightweight ontology repository. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems - AAMAS '03* (2003), p. 632.

[205] PARR, C. S., GURALNICK, R., CELLINESE, N., AND PAGE, R. D. M. Evolutionary informatics: Unifying knowledge about the diversity of life. *Trends in Ecology and Evolution 27*, 2 (2012), 94–103.

[206] PARR, C. S., SCHULZ, K. S., HAMMOCK, J., WILSON, N., LEARY, P., RICE, J., AND CORRIGAN, JR., R. J. TraitBank: Practical semantics for organism attribute data. *Semantic Web 7*, 6 (2016), 577–588.

[207] PARR, C. S., WILSON, N., LEARY, P., SCHULZ, K. S., LANS, K., WALLEY, L., HAMMOCK, J. A., GODDARD, A., RICE, J., STUDER, M., HOLMES, J. T. G., AND CORRIGAN, JR., R. J. The Encyclopedia of Life v2: Providing Global Access to Knowledge About Life on Earth. *Biodiversity Data Journal 2* (2014).

[208] PATTERSON, D. J., COOPER, J., KIRK, P. M., PYLE, R. L., AND REMSEN, D. P. Names are key to the big new biology. *Trends in Ecology and Evolution 25*, 12 (2010), 686–691.

[209] PATTERSON, D. J., REMSEN, D., MARINO, W. A., AND NORTON, C. Taxonomic Indexing—Extending the Role of Taxonomy. *Systematic Biology 55*, 3 (2006), 367–373.

[210] PEFFERS, K., ROTHENBERGER, M., TUUNANEN, T., AND VAEZI, R. Design Science Research Evaluation. In *Design Science Research in Information Systems. Advances in Theory and Practice: 7th International Conference, DESRIST 2012, Las Vegas, NV, USA, May 14-15, 2012. Proceedings* (2012), K. Peffers, M. Rothenberger, and B. Kuechler, Eds., Springer-Verlag, pp. 398–410.

[211] PEFFERS, K., TUUNANEN, T., ROTHENBERGER, M. A., AND CHATTERJEE, S. A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems 24*, 3 (2007), 45–77.

[212] PEREIRA, R., RICHARDS, K., HOBERN, D., HYAM, R., BELBIN, L., AND BLUM, S. TDWG life sciences identifiers (LSID) applicability statement. TDWG draft standard, Biodiversity Information Standards (TDWG), 2009. `http://www.tdwg.org/standards/150`.

[213] PHIPPS, J., AND HILLMANN, D. I. The Open Metadata Registry: An Update. *Bulletin of the Association for Information Science and Technology 37*, 4 (2011), 35–37.

[214] PULLAN, M. R., WATSON, M. F., KENNEDY, J. B., RAGUENAUD, C., AND HYAM, R. The Prometheus Taxonomic Model: A Practical Approach to Representing Multiple Classifications. *Taxon 49*, 1 (2000), 55–75.

[215] PYLE, R. L., AND MICHEL, E. ZooBank: Developing a nomenclatural tool for unifying 250 years of biological information. *Zootaxa*, 1950 (2008), 39–50.

[216] QING, Z. Building a platform for linked open thesauri. *Library Hi Tech 31*, 4 (2013), 620–637.

[217] QU, Y., AND CHENG, G. Falcons Concept Search: A Practical Search Engine for Web Ontologies. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 41*, 4 (2011), 810–816.

[218] REES, T. Taxamatch, an algorithm for near ('fuzzy') matching of scientific names in taxonomic databases. *PLoS ONE 9*, 9 (2014).

[219] REMSEN, D. The use and limits of scientific names in biological informatics. *ZooKeys 550* (2016), 207–223.

[220] REMSEN, D. P., DÖRING, M., AND ROBERTSON, T. GBIF GNA profile reference guide for Darwin Core Archive, core terms and extensions. Tech. rep., Global Biodiversity Information Facility (GBIF), Copenhagen, Denmark, 2011. version 1.2, released on 1 April 2011.

[221] RICHARDS, K., WHITE, R., NICOLSON, N., AND PYLE, R. A beginner's guide to persistent identifiers. Tech. rep., Global Biodiversity Information Facility (GBIF), Copenhagen, Denmark, 2011. version 1.0, released on 9 February 2011.

[222] RILEY, J. Understanding Metadata. Primer, National Information Standards Organization, 2017.

[223] ROSKOV, Y., ABUCAY, L., ORRELL, T., NICOLSON, D., FLANN, C., BAILLY, N., KIRK, P., BOURGOIN, T., DEWALT, R., DECOCK, W., AND DE WEVER, A. Species 2000 & ITIS Catalogue of Life. Digital resource, Species 2000: Naturalis, Leiden, Netherlands, 2017. 27th February 2017, `http://www.catalogueoflife.org/col`.

[224] RUEDA, C., BERMUDEZ, L., AND FREDERICKS, J. The MMI Ontology Registry and Repository: A Portal for Marine Metadata Interoperability. In *Proceedings of the Oceans Conference 2009, Biloxi, Mississippi, USA, 26–29 October* (2009), IEEE, pp. 1854–1859.

[225] SARKAR, I. N. Biodiversity informatics: Organizing and linking information across the spectrum of life. *Briefings in Bioinformatics 8*, 5 (2007), 347–357.

[226] SCHANDL, T., AND BLUMAUER, A. PoolParty: SKOS thesaurus management utilizing linked data. In *The Semantic Web: Research and Applications: 7th Extended Semantic Web Conference, ESWC 2010, Heraklion, Crete, Greece, May 30 – June 3, 2010, Proceedings, Part II* (2010), L. Aroyo, G. Antoniou, E. Hyvönen, A. ten Teije, H. Stuckenschmidt, L. Cabral, and T. Tudorache, Eds., Springer-Verlag, pp. 421–425.

[227] SCHREIBER, A. T., DUBBELDAM, B., WIELEMAKER, J., AND WIELINGA, B. Ontology-based photo annotation. *IEEE Intelligent Systems and Their Applications 16*, 3 (2001), 66–74.

[228] SCHULZ, S., STENZHORN, H., AND BOEKER, M. The ontology of biological taxa. *Bioinformatics 24*, 13 (2008), 313–321.

[229] SEGERS, H., DE SMET, W. H., FISCHER, C., FONTANETO, D., MICHALOUDI, E., WALLACE, R. L., AND JERSABEK, C. D. Towards a List of Available Names in Zoology, partim Phylum Rotifera. *Zootaxa 3179* (2012), 61–68.

[230] ŠEVČENKO, M. Online Presentation of an Upper Ontology. In *Proceedings of Znalosti 2003, Ostrava, Czech Republic, February 19-21* (2003), pp. 153–162.

[231] SHAW, R., RABINOWITZ, A., GOLDEN, P., AND KANSA, E. A sharing-oriented design strategy for networked knowledge organization systems. *International Journal on Digital Libraries 17*, 1 (2016), 49–61.

[232] SHIRI, A. Thesauri: Introduction and Recent Developments. In *Powering search: The role of thesauri in new information environments*. Information Today, 2012, pp. 1–41.

[233] SHVAIKO, P., AND EUZENAT, J. Ontology Matching: State of the Art and Future Challenges. *IEEE Transactions on Knowledge and Data Engineering 25*, 1 (2013), 158–176.

[234] SLIMANI, T. Semantic Annotation: The Mainstay of Semantic Web. *International Journal of Computer Applications Technology and Research 2*, 6 (2013), 763–770.

[235] SMITH, B., ASHBURNER, M., ROSSE, C., BARD, J., BUG, W., CEUSTERS, W., GOLDBERG, L. J., EILBECK, K., IRELAND, A., MUNGALL, C. J., OBI CONSORTIUM, LEONTIS, N., ROCCA-SERRA, P., RUTTENBERG, A., SANSONE, S.-A., SCHEUERMANN, R. H., SHAH, N., WHETZEL, P. L., AND LEWIS, S. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology 25*, 11 (2007), 1251–1255.

[236] SOERGEL, D. The Arts and Architecture Thesaurus (AAT): A Critical Appraisal. *Visual Resources 10*, 4 (1995), 369–400.

[237] SOLOMOU, G., AND PAPATHEODOROU, T. The use of SKOS vocabularies in digital repositories: The DSpace case. In *ICSC 2010: 2010 IEEE Fourth International Conference on Semantic Computing, Proceedings, Pittsburgh, Pennsylvania, USA, September 22-24* (2010), IEEE, pp. 542–547.

[238] SOUZA, R. R., TUDHOPE, D., AND ALMEIDA, M. B. The KOS spectra A tentative typology of knowledge organization systems. In *Paradigms and conceptual systems in knowledge organization: Proceedings of the Eleventh International ISKO Conference, Rome, Italy, 23-26 February* (2010), C. Gnoli and F. Mazzocchi, Eds., Ergon-Verlag, pp. 122–128.

[239] STAAB, S., MAEDCHE, A., AND HANDSCHUH, S. An annotation framework for the semantic Web. In *Proceedings of the First Workshop on Multimedia Annotation, Tokyo, Japan* (2001).

[240] STAAB, S., AND STUDER, R., Eds. *Handbook on Ontologies*. International Handbooks on Information Systems. Springer-Verlag, Berlin Heidelberg, 2004.

[241] STOJANOVIC, L. *Methods and Tools for Ontology Evolution*. PhD thesis, University of Karlsruhe, Germany, Aug. 2004.

[242] STOJANOVIC, L., MAEDCHE, A., MOTIK, B., AND STOJANOVIC, N. User-driven Ontology Evolution Management. In *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web: 13th International Conference, EKAW 2002, Sigüenza, Spain, October 1–4, 2002, Proceedings* (2002), A. Gómez-Pérez and V. R. Benjamins, Eds., Springer-Verlag, pp. 285–300.

[243] STUCKENSCHMIDT, H., AND KLEIN, M. Integrity and change in modular ontologies. In *IJCAI'03: Proceedings of the 18th international joint conference on Artificial intelligence* (2003), Morgan Kaufmann Publishers, pp. 900–905.

[244] STUDER, R., BENJAMINS, V. R., AND FENSEL, D. Knowledge engineering: Principles and methods. *Data & Knowledge Engineering 25*, 1-2 (1998), 161–197.

[245] 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 2008, Berlin, Germany, September 22-26* (2008), Dublin Core Metadata Initiative, pp. 25–33.

[246] SUNAGAWA, E., KOZAKI, K., KITAMURA, Y., AND MIZOGUCHI, R. An environment for distributed ontology development based on dependency management. In *The Semantic Web - ISWC 2003: Second International Semantic Web Conference, Sanibel Island, FL, USA, October 20-23, 2003, Proceedings* (2003), D. Fensel, K. Sycara, and J. Mylopoulos, Eds., Springer-Verlag, pp. 453–468.

[247] SUOMINEN, O., HYVÖNEN, E., VILJANEN, K., AND HUKKA, E. HealthFinland-A national semantic publishing network and portal for health information. *Journal of Web Semantics 7*, 4 (2009), 287–297.

[248] SUOMINEN, O., JOHANSSON, A., YLIKOTILA, H., TUOMINEN, J., AND HYVÖNEN, E. Vocabulary Services Based on SPARQL Endpoints: ONKI Light on SPARQL. In *Poster proceedings of the 18th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2012), Galway, Ireland, October 8-12* (2012).

[249] SUOMINEN, O., PESSALA, S., TUOMINEN, J., LAPPALAINEN, M., NYKYRI, S., YLIKOTILA, H., FROSTERUS, M., AND HYVÖNEN, E. Deploying national ontology services: From ONKI to Finto. In *Proceedings of the Industry Track at the International Semantic Web Conference 2014, Riva del Garda, Italy, October 19-23* (2014), A. Polleres, A. Garcia, and R. Benjamins, Eds., CEUR Workshop Proceedings.

[250] TAM, A. M., AND LEUNG, C. H. C. Structured Natural-Language Description for Semantic Content Retrieval. *Journal of the American Society for Information Science and Technology 52*, 11 (2001), 930–937.

[251] TAXONOMIC NAMES AND CONCEPTS INTEREST GROUP. Taxonomic concept transfer schema (TCS). TDWG current (2005) standard, Biodiversity Information Standards (TDWG), 2006. [http://www.tdwg.org/standards/117.

[252] TENNIS, J. T., AND SUTTON, S. A. Extending the Simple Knowledge Organization System for Concept Management in Vocabulary Development Applications. *Journal of the American Society for Information Science and Technology 59*, 1 (2008), 25–37.

[253] THESSEN, A. E., CUI, H., AND MOZZHERIN, D. Applications of natural language processing in biodiversity science. *Advances in Bioinformatics 2012* (2012).

[254] THOMAS, E., PAN, J. Z., AND SLEEMAN, D. ONTOSEARCH2: Searching ontologies semantically. In *Proceedings of the OWLED 2007 Workshop on OWL: Experiences and Directions, Innsbruck, Austria, June 6-7* (2007), C. Golbreich, A. Kalyanpur, and B. Parsia, Eds., CEUR Workshop Proceedings.

[255] TOPQUADRANT. Controlled vocabularies, taxonomies, and thesauruses (and ontologies). Whitepaper, TopQuadrant, 2013.

[256] TUDHOPE, D., ALANI, H., AND JONES, C. Augmenting thesaurus relationships: Possibilities for retrieval. *Journal of Digital Information 1*, 8 (2001).

[257] TUDHOPE, D., AND BINDING, C. Toward Terminology Services: Experiences with a Pilot Web Service Thesaurus Browser. *Bulletin of the American Society for Information Science and Technology 32*, 5 (2006), 6–9.

[258] TUOMINEN, J., SALONOJA, M., VILJANEN, K., AND HYVÖNEN, E. A user interface for ontology repositories. In *Proceedings of the 1st Workshop on Ontology Repositories and Editors for the Semantic Web, Hersonissos, Crete, Greece, May 31* (2010), M. d'Aquin, A. García Castro, C. Lange, and K. Viljanen, Eds., CEUR Workshop Proceedings, pp. 17–21.

[259] TUOMINEN, J., AND VILJANEN, K. Ontology services: results of the ONKI user inquiry, presentation at the FinnONTO 2.0 project board meeting 13.12.2011, 2011.

[260] USCHOLD, M., AND GRUNINGER, M. Ontologies: Principles, methods and applications. *Knowledge Engineering Review 11*, 2 (1996), 93–136.

[261] VAN ASSEM, M., MALAISÉ, V., MILES, A., AND SCHREIBER, G. A method to convert thesauri to SKOS. In *The Semantic Web: Research and Applications: 3rd European Semantic Web Conference, ESWC 2006, Budva, Montenegro, June 11-14, 2006, Proceedings* (2006), Y. Sure and J. Domingue, Eds., Springer-Verlag, pp. 95–109.

[262] VAN ASSEM, M., MENKEN, M. R., SCHREIBER, G., WIELEMAKER, J., AND WIELINGA, B. A Method for Converting Thesauri to RDF/OWL. In *The Semantic Web – ISWC 2004: Third International Semantic Web Conference, Hiroshima, Japan, November 7-11, 2004, Proceedings* (2004), S. A. McIlraith, D. Plexousakis, and F. van Harmelen, Eds., Springer-Verlag, pp. 17–31.

[263] VAN DEN BOSCH, A., BOGERS, T., AND DE KUNDER, M. Estimating search engine index size variability: a 9-year longitudinal study. *Scientometrics 107*, 2 (2016), 839–856.

[264] VAN DER MEIJ, L., ISAAC, A., AND ZINN, C. A web-based repository service for vocabularies and alignments in the cultural heritage domain. In *The Semantic Web: Research and Applications: 7th Extended Semantic Web Conference, ESWC 2010, Heraklion, Crete, Greece, May 30 –June 3, 2010, Proceedings, Part I* (2010), L. Aroyo, G. Antoniou, E. Hyvönen, A. ten Teije, H. Stuckenschmidt, L. Cabral, and T. Tudorache, Eds., Springer-Verlag, pp. 394–409.

[265] VAN OSSENBRUGGEN, J., AMIN, A., AND HILDEBRAND, M. Why evaluating semantic web applications is difficult. In *Proceedings of the Fifth International Workshop on Semantic Web User Interaction (SWUI 2008), collocated with CHI 2008, Florence, Italy, April 5* (2008), D. Degler, M. Schraefel, J. Golbeck, A. Bernstein, and L. Rutledge, Eds., CEUR Workshop Proceedings.

[266] VANDEN BERGHE, E., CORO, G., BAILLY, N., FIORELLATO, F., ALDEMITA, C., ELLENBROEK, A., AND PAGANO, P. Retrieving taxa names from large biodiversity data collections using a flexible matching workflow. *Ecological Informatics 28* (2015), 29–41.

[267] VANDENBUSSCHE, P.-Y., ATEMEZING, G. A., POVEDA-VILLALÓN, M., AND VATANT, B. Linked Open Vocabularies (LOV): a gateway to reusable semantic vocabularies on the Web. *Semantic Web*, In press (2016).

[268] VILJANEN, K., HYVÖNEN, E., MÄKELÄ, E., SUOMINEN, O., AND TUOMINEN, J. Mash-up ontology services for the semantic web. In *Demo track at the European Semantic Web Conference ESWC 2007, Innsbruck, Austria, June 4* (2007).

[269] VIZINE-GOETZ, D., HOUGHTON, A., AND CHILDRESS, E. Web Services for Controlled Vocabularies. *Bulletin of the American Society for Information Science and Technology 32*, 5 (2006), 9–12.

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

[271] WALLS, R. L., DECK, J., GURALNICK, R., BASKAUF, S., BEAMAN, R., BLUM, S., BOWERS, S., BUTTIGIEG, P. L., DAVIES, N., ENDRESEN, D., GANDOLFO, M. A., HANNER, R., JANNING, A., KRISHTALKA, L., MATSUNAGA, A., MIDFORD, P., MORRISON, N., TUAMA, É. Ó., SCHILDHAUER, M., SMITH, B., STUCKY, B. J., THOMER, A., WIECZOREK, J., WHITACRE, J., AND WOOLEY, J. Semantics in Support of Biodiversity Knowledge Discovery: An Introduction to the Biological Collections Ontology and Related Ontologies. *PLoS ONE 9*, 3 (2014).

[272] WANG, S., SCHLOBACH, S., AND KLEIN, M. Concept drift and how to identify it. *Journal of Web Semantics 9*, 3 (2011), 247–265.

[273] WANG, X., ALMEIDA, J. S., AND OLIVEIRA, A. L. Ontology design principles and normalization techniques in the web. In *Data Integration in the Life Sciences: 5th International Workshop, DILS 2008, Evry, France, June 25-27, 2008, Proceedings* (2008), A. Bairoch, S. Cohen-Boulakia, and C. Froidevaux, Eds., Springer-Verlag, pp. 28–43.

[274] WANG, Y.-C., VANDENDORPE, J., AND EVENS, M. Relational thesauri in information retrieval. *Journal of the American Society for Information Science 36*, 1 (1985), 15–27.

[275] WHETZEL, P. L., NOY, N. F., SHAH, N. H., ALEXANDER, P. R., NYULAS, C., TUDORACHE, T., AND MUSEN, M. A. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Research 39*, Web Server issue (2011), W541–W545.

[276] WHITE, H., WILLIS, C., AND GREENBERG, J. HIVEing: The Effect of a Semantic Web technology on Inter-Indexer Consistancy. *Journal of Documentation 70*, 3 (2014), 307–329.

[277] WIECZOREK, J., BLOOM, D., GURALNICK, R., BLUM, S., DÖRING, M., GIOVANNI, R., ROBERTSON, T., AND VIEGLAIS, D. Darwin Core: An

evolving community-developed biodiversity data standard. *PLoS ONE 7*, 1 (2012).

[278] WIELINGA, B. J., SCHREIBER, A. T., WIELEMAKER, J., AND SANDBERG, J. A. C. From thesaurus to ontology. In *Proceedings of the 1st International Conference on Knowledge Capture: K-CAP 2001, Victoria, Canada, October 21-23* (2001), ACM, pp. 194–201.

[279] XIANG, Z., MUNGALL, C., RUTTENBERG, A., AND HE, Y. Ontobee: A linked data server and browser for ontology terms. In *Proceedings of the 2nd International Conference on Biomedical Ontology, Buffalo, NY, USA, July 26-30* (2011), O. Bodenreider, M. E. Martone, and A. Ruttenberg, Eds., CEUR Workshop Proceedings.

[280] YTOW, N., MORSE, D. R., AND ROBERTS, D. M. Nomencurator: a nomenclatural history model to handle multiple taxonomic views. *Biological Journal of the Linnean Society 73* (2001), 81–98.

[281] ZABLITH, F., ANTONIOU, G., D'AQUIN, M., FLOURIS, G., KONDYLAKIS, H., MOTTA, E., PLEXOUSAKIS, D., AND SABOU, M. Ontology evolution: a process-centric survey. *The Knowledge Engineering Review 30*, 1 (2015), 45–75.

[282] ZAPILKO, B., SCHAIBLE, J., MAYR, P., AND MATHIAK, B. TheSoz: A SKOS representation of the thesaurus for the social sciences. *Semantic Web 4*, 3 (2013), 257–263.

[283] ZENG, M. L., AND CHAN, L. M. Trends and issues in establishing interoperability among knowledge organization systems. *Journal of the American Society for Information Science and Technology 55*, 5 (2004), 377–395.

[284] ZENG, M. L., AND HODGE, G. Developing a Dublin Core Application Profile for the Knowledge Organization Systems (KOS) Resources. *Bulletin of the American Society for Information Science & Technology 37*, 4 (2011), 30–34.

[285] ZERMOGLIO, P. F., GURALNICK, R. P., AND WIECZOREK, J. R. A standardized reference data set for vertebrate taxon name resolution. *PLoS ONE 11*, 1 (2016).

[286] ZHAO, L., AND ICHISE, R. Ontology Integration for Linked Data. *Journal on Data Semantics 3*, 4 (2014), 237–254.

Publication I

Kim Viljanen, Jouni Tuominen, and Eero Hyvönen. Ontology Libraries for Production Use: The Finnish Ontology Library Service ONKI. In *The Semantic Web: Research and Applications: 6th European Semantic Web Conference, ESWC 2009, Heraklion, Crete, Greece, May 31–June 4, 2009, Proceedings*, Lora Aroyo, Paolo Traverso, Fabio Ciravegna, Philipp Cimiano, Tom Heath, Eero Hyvönen, Riichiro Mizoguchi, Eyal Oren, Marta Sabou, and Elena Simperl (editors), Lecture Notes in Computer Science, volume 5554, pages 781–795, ISBN 978-3-642-02120-6, Springer-Verlag, June 2009.

Ontology Libraries for Production Use: The Finnish Ontology Library Service ONKI

Kim Viljanen, Jouni Tuominen, and Eero Hyvönen

Semantic Computing Research Group (SeCo)
Helsinki University of Technology (TKK) and University of Helsinki
firstname.lastname@tkk.fi
http://www.seco.tkk.fi/

Abstract. This paper discusses problems of creating and using ontology library services in production use. One approach to a solution is presented with an online implementation—the Finnish Ontology Library Service ONKI— that is in pilot use on a national level in Finland. ONKI contributes to previous research on ontology libraries in many ways: First, mashup and web service support with various tools is provided for cost-efficient utilization of ontologies in indexing and search applications. Second, services covering the different phases of the ontology life cycle are provided. Third, the services are provided and used in real world applications on a national scale. Fourth, the ontology framework is being developed by a collaborative effort by organizations representing different application domains, such as health, culture, and business.

1 Introduction

The Semantic Web[1] is based on ontologies [1,2,3]. With the help of ontologies, the content and services on the Web can be described with metadata in an explicit, machine "understandable" way which enables, for example, interoperability on a semantic level and intelligent semantic searching and browsing of heterogeneous distributed content in semantic portals [4,5,6]. Utilizing ontologies, including thesauri and other vocabularies, in new and existing applications requires efficient tools for finding, managing, searching and browsing ontologies. *Ontology library systems*[2] offer functions for managing, adapting and standardizing groups of ontologies, for indexing content with ontologies, and for utilizing ontologies in applications [6,7,8,9].

This paper discusses the requirements for ontology library systems from a practical viewpoint, and presents an approach for building such a service. As a concrete result of the research and a case study, the national Finnish Ontology Library Service ONKI[3] is presented. ONKI is a major objective of the National

[1] http://www.w3.org/2001/sw

[2] Ontology library systems are referred to in the literature with terms "ontology servers" and "ontology services", too. We use the term "ontology library systems" to encompass all these meanings.

[3] http://www.yso.fi/

L. Aroyo et al. (Eds.): ESWC 2009, LNCS 5554, pp. 781–795, 2009.

Semantic Web Ontology project FinnONTO (2003–2010)[4] which aims at developing a Semantic Web ontology infrastructure on a national level in Finland [6]. The consortium behind the initiative represents a wide spectrum of functions of the society, including libraries, health organizations, cultural institutions, government, media, and education.

In the following, we first set requirements for an ontology library system in terms of services at different phases of ontology development and usage. After this, the ONKI system and its services are presented along the same phases, and the implementation is discussed. The system has been used in several case applications that are briefly surveyed next. In conclusion, related work is discussed and contributions of ONKI summarized.

2 Requirements for an Ontology Library Service

Requirements for ontology library services can be identified by analyzing the life cycle of an ontology. Based on literature [8,7] and our own work, we identify following parts to be typical in a life cycle of an ontology.

1) Designing the ontologies. The first step in the life cycle of an ontology is to design the structure and modelling principles of the ontology based on analysis of the subject domain and by identifying the business and application problems the ontology is intended to solve. The foundational classes, properties and instances of the ontology are created. Sometimes the ontology may also be based on existing vocabularies such as a thesaurus, which are then "ontologised". The main actors of this phase are workgroups and individual ontologists. Ontology libraries should provide support e.g. for collaborative ontology editing, reuse and alignment [10]. *2) Populating the ontologies.* Ontologies may consist of huge amounts of instances such as people, organisations and places. Populating and maintaining the information can either be a one-time effort or constantly continuing process even after the ontology has been published. Populating may be e.g. a community-based distributed effort or based on utilizing existing registries and sources as input. Ontology libraries should support such content collection and updating processes. *3) Publishing the ontologies.* When an ontology has been created, methods for publishing and promoting it are needed to ensure that the ontology is actively used to achieve the benefits of creating the ontology in the first place. In addition to provide such publishing and promoting mechanisms, the ontology library should also have mechanisms - both manual and automatic - for ensuring the quality of the ontologies to be published. The main actors of this phase are the ontology owners. *4) Finding, comparing and commiting to ontologies.* When considering using ontologies for some purpose, the finding of suitable ontologies require support from the ontology library. Typical users of this phase are information architects. *5) Ontology based semantic application creation.* Learning, evaluating and implementating ontology services to applications require functionalities from the ontology library service for making this

[4] Our work is funded by the National Funding Agency for Technology and Innovation (Tekes) and a consortium of 38 companies and public organizations. http://www.seco.tkk.fi/projects/finnonto/

process as fluent and easy as possible - and to support the wide usage of ontologies in applications [11,12]. Typical users of this phase are software architects. *6) Ontology based semantic content creation.* Ontologies are mostly used for describing and indexing content semantically. Ontology libraries should support the work of content indexers by providing efficient tools and services for e.g. browsing ontologies, finding and fetching concepts for annotating purposes [9] or automatically indexing documents [13]. *7) Ontology-based end-user applications.* Ontological content search, semantic browsing, semantic portals are examples of typical ways to provide the end-user with benefits from using ontologies in an applications (see e.g. [4,5,6,14]). Ontology libraries should support creating such end-user applications by providing services for the application builders. Ontology libraries may also provide services directly to the end-users such as the possibility to learn about some domain with the help of ontologies.

3 Finnish Ontology Library Service ONKI

The Finnish Ontology Library Service ONKI is a pilot system for addressing the requirements of an ontology library service on a national scale, but with the special focus on ontology publishing and using them in content indexing, and information retrieval through both user and application interfaces [14,6]. ONKI contains currently over 40 ontologies from various domain areas (see Table 1). Most of the ontologies are freely available to anybody to test and use in their applications.

3.1 Designing the Ontologies, Populating the Ontologies

Ontologies developed within the ONKI framework are mostly created with the Protégé[5] editor. Version management of the ontology files is done with Subversion[6]. In many cases, the ontologies are based on an existing thesaurus or other content which have first been transformed with an custom-made program to OWL and then been refined ontologically by the ontologist using Protégé and by aligning the ontologies with the Finnish Upper Ontology YSO [6]. Populating the ontologies have been done either with Protégé by the ontologist(s) or collaboratively with the browser-based annotation editor SAHA [15]. For many ontologies, populating have been done with custom-made programs.

Quality of the ontologies is controlled using three ways: gate keeping, quality requirements and training. Gate keeping is practised by selecting only trusted participants in the ontology creation and publishing which include both companies, governmental and non-governmental organizations. Typically the main author of any single ontology in ONKI is the leading authority in Finland of the respective domain. Quality requirements enforced in ONKI cover ontology presentation and ontology creation process issues. The ontology should be presented using some RDF-based ontology language, such as SKOS, OWL or RDF

[5] http://protege.stanford.edu
[6] http://subversion.tigris.org/

Table 1. A selection of ontologies currently available in ONKI (The amount of concepts consists classes and/or instances, depending on the ontology.)

Ontology	Concepts	Format	Public?
Upper and Holistic Ontologies			
Holistic Collaborative Finnish Ontology KOKO	ca. 30,000	OWL	yes
General Finnish Upper Ontology YSO	20,649	OWL	yes
General Finnish Thesaurus YSA	26,633	SKOS	yes
Wordnet	ca. 230,000	SKOS	yes
Cultural Ontologies			
Ontology for Museum Domain MAO	6,775	OWL	yes
Ontology of Applied Arts TAO	29,940	OWL	yes
Finnish Ontology of Photography VALO	22,596	OWL	yes
Ontology for music MUSO	21,650	OWL	yes
Art and Iconography classification Iconclass	26,636	SKOS	yes
Kaunokki thesaurus for fictive literature	4,373	SKOS	yes
Music thesaurus MUSA	931	SKOS	yes
Art & Architecture Thesaurus AAT	27,992	OWL	no
Agriforest and Natural Science Ontologies			
Agriforest Ontology AFO	26,612	OWL	yes
Ontology of Birds AVIO	11,161	SKOS	no
Ontology of Mammals MAMO	6,059	SKOS	no
Health Ontologies			
Medical Subject Headings MeSH	24,355	SKOS	yes
European Multilingual Thesaurus on Health Promotion HPMULTI	1,271	SKOS	yes
Business Ontologies, Governmental Ontologies			
Seafaring thesaurus MESA	1,448	SKOS	yes
United Nations Standard Products and Services Code UNSPSC	20,794	SKOS	no
Finnish Governmental Thesaurus VNAS	6,342	SKOS	yes
Instance Ontologies			
Finnish Geo-ontology SUO	ca. 800,000	OWL	yes
Finnish Time-Location Ontology SAPO	1,102	OWL	yes
Getty Thesaurus of Geographic Names TGN (exluding USA)	142,990	OWL	no
Getty Union List of Artist Names ULAN	ca. 100,000	OWL	no

Schema. The consistency of the ontologies is checked including syntax checking (valid RDF) and conceptual checking manually by the ontologists. One of the most important ways to enforce the quality is that the ontologies in ONKI library are created using a common development process and modelling idea which are supervised by the core YSO developer team [6]. This promotes using compatible development processes in all other ontologies also and thus provides a more compatible collection of ontologies as a result. If possible, ontologies are aligned with a common upper ontology, the Finnish Upper Ontology YSO. This alignment to YSO adds value to YSO, the ontology at hand and the ONKI Library as a whole, because each additional alignment adds new possibilities to find concept relations. To spread good practices and knowledge about the modelling methods used in YSO and other relevant ontologies, training is provided to ontology developers. To enforce the reuse of ontologies, the license of the published ontologies should allow publishing, using and redeveloping the ontology as

freely as possible. The default license used for ONKI ontologies is the Creative Commons license[7].

3.2 Publishing the Ontologies

Publishing an ontology in ONKI typically contains the following phases: First, the ontology to be published is added to the Subversion repository and the needed configurations for the ontology are created [16]. Second, a URI normalization for the ontology to be published is done where the original URIs are transformed to persistent numeric URIs (PURIs). Instead of (typically) human readable URIs we propose that URIs should not contain any reference to human languages to avoid unnecessary needs for changing the concept URIs e.g. when translating the ontology to some other language[8]. For example, instead using the URI "myonto:semanticweb" we propose using the URI "myonto:p12345". Third, if the ontology is published part of the KOKO ontology, the automatic updating of KOKO takes place. Finally, the ontology is added to ONKI and made available via different services such as human user interfaces and machine APIs.

If the ontology is maintained in some external system it can be published using ONKI by establishing a publishing pipeline from the external system to ONKI. This method has been used e.g. in publishing the General Finnish Thesaurus YSA, maintained by the National Library of Finland [16]. The thesaurus is fetched each night from the National Library's server using the MARCXML format[9]. The content is then transformed to SKOS and finally published in ONKI. ONKI provides also an upload functionality "Your ONKI" for publishing SKOS (or other) ontologies in the library. When an ontology has been uploaded, it is moderated by the server administration and if the content is suitable, it will be added to the library. ONKI quality requirements presented earlier are recommended also for Your ONKI submissions.

3.3 Ontology Discovery, Ontology Library Service Evaluation

To support finding, evaluating and choosing an ontology for specific purposes each ontology is described with metadata including title, description, classification, version information and available access methods. Depicted in Figure 1 is the main user interface with the list of available ontologies, which also shows the available access methods for each ontology. Access methods are described later, but include e.g. the possibility to browse and search the ontologies. The ontologies can be described and documented in a wiki, part of the ONKI system.

The list of ontologies is available in RDF for machine usage. When allowed by the publishing license, current and previous versions of ontologies are available for downloading.

[7] http://creativecommons.org/
[8] The idea of stable URIs is also discussed in http://www.w3.org/Provider/Style/URI
[9] http://www.loc.gov/standards/marcxml/

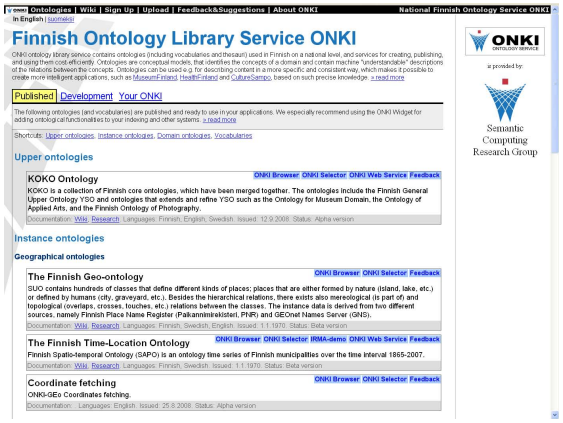


Fig. 1. List of ontologies in ONKI. Each ontology contains the links to the available access methods such as the ontology-specific browser.

3.4 Ontology-Based Semantic Content Creation

For users creating semantic content, e.g. by describing resources by using ontological concepts, ONKI service provides the ONKI Selector, depicted in Figure 2 [9]. With ONKI Selector content creators can find suitable concepts for their annotation tasks. When the ONKI Selector is integrated into a HTML input field, the field turns into a semantic autocompletion search interface. When typing search string into the input field, the matching concepts are returned as a hit list. Desired concepts can be selected from the list and added to the content creation application. Depending on the use case, concept's URI, label or both of them can be fetched to the application.

In combination with the ONKI Selector, domain-specific ONKI browsers can be used to browse the ontologies when searching for suitable concepts. The browsers have a "Fetch Concept" button which returns the selected concept into the content creation application. ONKI SKOS Browser[16] is an ontology

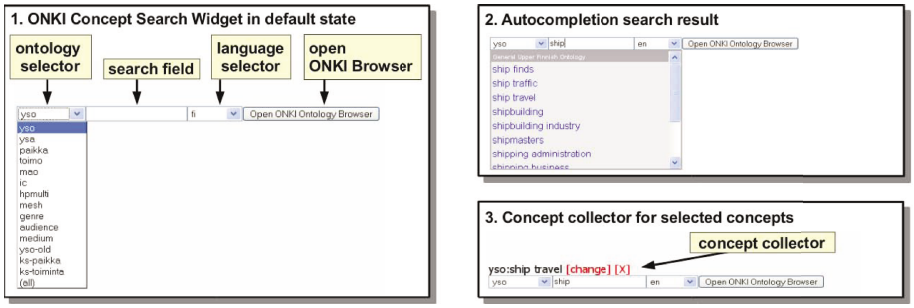


Fig. 2. ONKI Selector

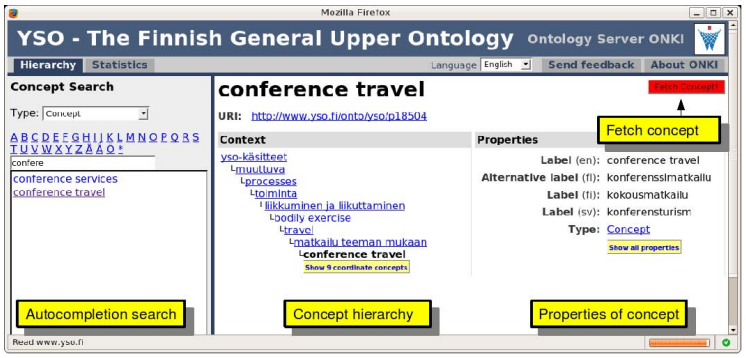


Fig. 3. ONKI SKOS Browser

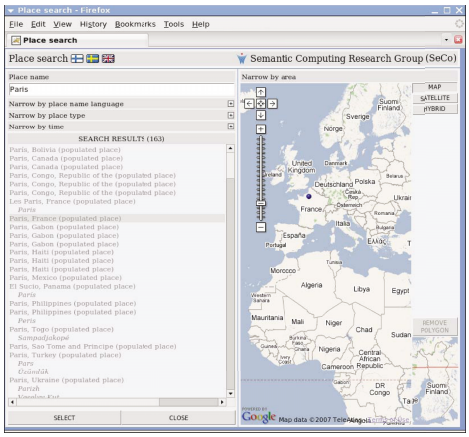


Fig. 4. ONKI Geo Browser

browser for thesaurus-like class ontologies. It supports visualizing and browsing of vocabularies conforming to SKOS recommendation, and also RDF(S) and OWL ontologies with additional configuration. ONKI SKOS Browser consists of three main components: 1) *concept search with semantic autocompletion*, 2) *concept hierarchy* and 3) *concept properties*, as depicted in Figure 3. ONKI Geo Browser [17] is used for accessing geographical instance data with a map interface, as depicted in Figure 4. It provides unambiguous place identifiers (URIs) and coordinates for arbitrary points or polygons to be used in content annotation. ONKI People [18] is used for browsing and searching ontologies of persons, organizations, and similar instance registries.

3.5 Ontology-Based End-User Applications

For ontology-based end-user applications ONKI service provides means for finding ontological concepts and using them, e.g., in information retrieval tasks.

Compared to a simple free text search field, the ONKI Selector aids user to find query concepts with autocompletion search and ontology browsers. The ONKI Selector is useful even if the application is not ontology-based. In that case the labels of the concepts can be used as query terms.

To increase the recall of the information retrieval tasks, ONKI Selector performs query expansion by ontological inference. The properties used for performing the query expansion can be configured separately for each ontology. In class ontologies, a concept is typically expanded to its subconcepts. The query expansion could also be based on partonomy, associative relationships or other relations between concepts. In geographical instance data, a place instance is expanded to places that have historically had overlapping regions with the place. Other possible query expansion methods include partonomy, places with shared regional borders etc.

3.6 Ontology-Based Semantic Application Creation

ONKI services may be integrated into semantic applications at the user interface level by using the ready-to-use user interface component ONKI Selector, domain-specific ONKI Browsers, and by using application programming interfaces (API). ONKI supports the software developer in using ONKI services with the help of helper applications such as the ONKI Selector Builder, depicted in Figure 5, which helps the developer to generate the JavaScript code needed for integrating ONKI Selector into web-based applications. When the desired configuration properties have been set in the ONKI Selector Builder, the resulting JavaScript code can be copied into the application being developed.

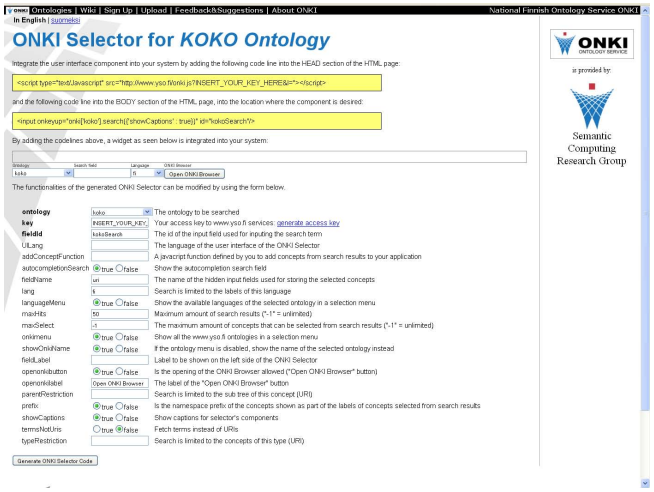


Fig. 5. ONKI Selector builder

ONKI API provides methods for accessing ontologies, e.g., for searching for concepts, getting metadata of an ontology and performing ontology-based query expansion. ONKI API is implemented as Web Service (SOAP) and JavaScript interface. ONKI API contains the following methods:

- *search(query, lang, maxHits, type, parent)* - for searching for the ontological concepts. Returns a list of matching concepts.
- *expandQuery(URI, lang, maxHits, type)* - for querying for the query expansion for a concept. Returns a list of concepts.
- *getLabel(URI, lang)* - for fetching a label for a given concept URI in a given language.
- *getAvailableLanguages()* - for querying for the supported languages of an ontology. Returns a list of language codes.
- *getAvailableTypeUris()* - for querying for the concept types (rdf:type relations) existing in the ontology. Returns a list of URIs.

Software developers can also utilize the RDF files of the ontologies published in the ONKI service. All concept and instance URIs are designed so that they function also as URLs. When the URI of a concept is accessed with a web browser, the relevant view is opened in the ONKI browser. This means that the URI itself acts as a functional link when added to a HTML page. In accordance to W3C[10], if the URI is accessed with an RDF aware system, the machine readable RDF presentation of the content is returned instead of the ONKI browser's HTML presentation.

4 Implementation and Usage Statistics

The ONKI service is constituted of a loosely coupled set of independent applications such as the ONKI SKOS and ONKI Geo servers which are combined by using a lightweight facade service made with an Apache web server, Apache rewrite rules and PHP scripts. Back-end ONKI applications conform also to the ONKI API described in section 3.6. Technologies used for implementing the various back-end applications include Java, Semantic Web Framework Jena[11], MySQL database, Lucene index[12], Direct Web Remoting DWR for AJAX functionalities[13], Varnish HTTP accelerator[14] and shell scripts and Subversion version management system. The ontologies are presented internally in various RDF formats, typically in SKOS or OWL. With the help of ontology-specific configurations, the ontologies are served to the user in a uniform way.

The ONKI is running as a pilot service publicly available on the web. It was officially launched in September 2008[15]. During year 2008 ONKI had ca. 36,000

[10] http://www.w3.org/TR/swbp-vocab-pub/
[11] http://jena.sourceforge.net
[12] http://lucene.apache.org
[13] http://www.directwebremoting.org
[14] http://varnish.projects.linpro.no
[15] http://www.youtube.com/watch?v=qG2YhK17ifs

Table 2. ONKI usage statistics for November 2008

Service	Hits
Human interfaces	
ONKI-SKOS Browser	89,346
ONKI Selector Widget	54,384
Persistent URI redirects	4,415
ONKI Selector builder	1,103
Web Service builder	1,403
ONKI-IRMA	495
ONKI-Geo Browser	203
ONKI-Geo coordinates Browser	168
Application interfaces	
Web service calls	87,388
Javascript calls	18,816
Total	257,721

unique visitors and ca. 104,000 visits. 91 organizations outside the research group have registered an access key for using the JavaScript and web service interfaces which of 25 have actually implemented a test application using the components. For an overview, table 2 presents the usage statistics of the main ONKI functionalities for a representative month[16]. The most popular ontologies are the Medical Subject Headings, the Finnish Upper Ontology YSO, the General Finnish Thesaurus YSA and the Ontology for Museum Domain MAO where each ontology got over ten thousand hits during the month.

5 Case Applications Using ONKI

5.1 Content Creation: HealthFinland, CultureSampo and Tilkut

HealthFinland and CultureSampo are two major pilot applications of the FinnONTO project [6]. They demonstrate the usage of Semantic Web technologies in the contexts of health promotion and cultural heritage. Both systems uses ONKI as the ontology server for indexing content especially with the the browser-based annotation editor SAHA[17] [15]. For example (Figure 6), the Finnish General Ontology YSO [6] has been added as a ONKI Selector component to SAHA for finding and fetching annotation concepts.

The web laboratory Owela[18] of the VTT Technical Research Centre of Finland has implemented a service for collecting and sharing text and image clips from the Web[19]. In the service one can organize the clips into folders and tag them with different categories. The ONKI Selector is used for tagging the clips.

[16] The hits represents user activities of the system. Outliers caused e.g. by web crawlers have been removed as far as possible.
[17] http://www.seco.tkk.fi/services/saha/
[18] http://owela.vtt.fi/
[19] http://owela.vtt.fi/tilkut

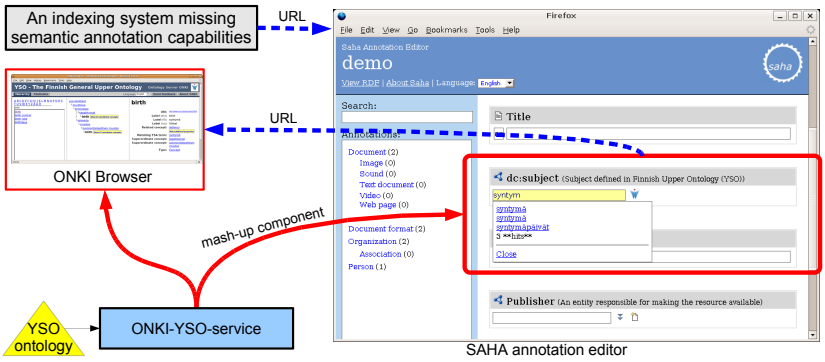


Fig. 6. The Finnish General Ontology connected to SAHA

5.2 Content Search: Kantapuu.fi and eViikki

Kantapuu.fi[20] is a web user interface for browsing and searching for collections of Finnish museums of forestry. The collection items are annotated with terms from General Finnish Thesaurus YSA, Thesaurus for Museum Domain MASA and Agriforest Thesaurus. Kantapuu.fi search page is a web form into which query strings are typed as free text. The query strings can be placed into specific fields, e.g. "keywords", "place of use" or "time of use". We have created a demonstration page containing a Kantapuu.fi's search form with integrated ONKI Selectors which can be used for selecting query terms to be used in the Kantapuu.fi search[21]. The ONKI Selector is used for finding terms from vocabularies of the ONKI service. The used vocabularies are the same as those used in the annotation process of the items, or actually their ontologized versions. To find suitable query terms user can utilize the autocompletion search functionality or the ONKI Browser. Thus, the user does not need to be familiar with the vocabularies used in the annotations of the items, as in the case of free text search. The ONKI Selector performs query expansion based on the selected query terms. So, for example a query term "animals" would return items annotated with term "cats". When the desired query terms are selected, the actual search to the Kantapuu.fi system can be executed.

The ONKI Selector Widget has also been integrated into the Viikki Science Library[22] reference database system eViikki[23]. eViikki is a search interface for the library's collections, which consist of scientific literature on agriforestry. The ONKI Selector is used for populating the "keywords" field of the search form of eViikki. The fetched concept labels are used in the information retrieval task. Query expansion is not performed currently.

[20] http://www.kantapuu.fi/
[21] http://www.yso.fi/lusto
[22] http://www.tiedekirjasto.helsinki.fi/english/
[23] http://www-db.helsinki.fi/eviikki/eviikkihaku.html

6 Related Work

Based on reviews on ontology library systems [7,8], the main focus in existing systems tends to be in supporting ontology development and not the runtime usage of ontologies such as indexing and ontology-based end-user applications. Although ONKI provides support for the whole ontology life cycle, a major contribution of ONKI is the support for indexing and other runtime needs.

The DAML Ontology Library[24] is a classic implementation of an ontology library. The ontologies can be accessed via different categories, such as metadata describing the ontologies such as keywords, Open Directory category[25] and submitting organization. Also information obtained from the ontologies such as names of classes and properties can be used for finding relevant ontologies. The main method for using the ontologies is to download them. In comparison, ONKI provides application support for e.g. adding ontologies as mash-up and web services to applications.

The Ontology Library Service ONKI provides methods for finding ontologies and concepts amongst the ontologies published in the centralized service, whereas Swoogle [19] and Watson [20] act as global Semantic Web search engines. They crawl the web and index the RDF files they find. Such search engines can be useful when searching for suitable ontologies to use in applications, providing an overview of ontologies of some domain published on the web. The ONKI Service is based on a different approach. It aims to be a community-based service that gathers together the users of ontologies providing them ready-to-use ontological functionalities which can be integrated into semantic applications.

Dameron et al. proposes that ontology services should be provided as Ontology Web Services (OWS) which could be used in applications for automatically find and use ontologies [12]. In ONKI we support the idea of providing application interfaces to the ontology library, but extend the idea to a higher abstraction level by providing also ready-to-use user interface components to avoid duplicated work by re-implementing user interface and visualization functionalities.

Faviki[26] is a semantic bookmarking service which uses Wikipedia's[27] term identifiers for tagging web pages. In comparison, ONKI is focusing on publishing ontologies and to support the creation of ontology-based indexing, content search and other applications.

Freebase[28] is a data repository on the web that aggregates information from many sources and provides a single topic and identifier for each logical entity, e.g. a person. One goal of ONKI is also to provide (optimally) single, shared identifiers for ontological concepts which can be used to aggregate distributed content repositories. Freebase is based on a bottom-up approach based on existing information e.g. in Wikipedia whereas ONKI ontologies are (typically) based on top-down analysis of a domain and its relevant concepts.

[24] http://www.daml.org/ontologies/
[25] http://www.dmoz.org
[26] http://www.faviki.com
[27] With the help of DBPedia, http://dpbedia.org
[28] http://www.freebase.com

7 Discussion

This paper discussed the requirements of an ontology library system to support the different phases of an ontology life cycle and related user needs for creation, publishing, maintaining and using ontologies. The Finnish national ontology library service ONKI addresses all phases of the ontology life cycle and contributes especially in providing support for 1) collaborative ontology publishing, 2) content indexing, and 3) information searching. The ontology services can be used in external legacy and other applications as ready-to-use functionalities. The new idea here is to support mash-up usage of ontologies in a way similar to Google Maps and other similar services. Our approach of providing an integrable autocompletion widget for external systems is the same as in [21].

The ONKI system supports syntactically, structurally and semantically heterogeneous content. RDF-based content representations such as RDF Schema, SKOS and OWL can be easily published by the ONKI SKOS server. SKOS generators for especially thesauri presentation formats such as MARCXML, various database schemas and text files has been implemented. Also multilingual content is supported. ONKI has been built for and tested with real world data consisting of ontologies, well-known thesauri and registries. The geographical ontology SUO contains over 800,000 places in Finland, which is published using the ONKI Geo server. The typical size of ontologies and thesauri published using the ONKI SKOS server is tens of thousands concept, e.g., YSO contains 20,600 concepts. For testing the scalability of ONKI SKOS, the Wordnet with 230,000 concepts has been successfully presented with the system. The ONKI services have been tested in various ways: the ONKI Selector as part of the SAHA editor in creating content for e.g. HealthFinland and CultureSampo and by external parties in their indexing and search applications. The ONKI Browser has been tested by expert users from e.g. the National Library of Finland.

To conclude, this full scale national ontology library service ONKI is novel and has not been done before. There are currently thousands of individual users and hundreds of organizations from different domains testing the ONKI system and using it in pilot applications. The National Library's commitment to ONKI means that a substantial part of public and private organizations in Finland begin to use ONKI – and most promisingly also the KOKO ontology – for indexing and search, but also for publishing and accessing their ontologies and thesauri – and to join the Semantic Web.

Future work include continuing observing how the ontology development, publishing and using community continues to build up around ONKI. We are most interested in seeing what kind of new applications will emerge based on ONKI and how well the concept of interlinked ontologies works in practice. DBPedia could be an interesting repository to be published in ONKI. Research topics include developing further methods for supporting community-based ontology development, managing changes in ontologies and utilizing change history and change propagation in e.g. inference and searching.

References

1. Gruber, T.R.: A translation approach to portable ontology spesifications. Knowledge Acquisition 5(2), 199–220 (1993)
2. Staab, S., Studer, R.: Handbook on ontologies. Springer, Heidelberg (2004)
3. Fensel, D.: Ontologies: Silver bullet for knowledge management and electronic commerce, 2nd edn. Springer, Heidelberg (2004)
4. Reynolds, D., Shabajee, P., Cayzer, S.: Semantic Information Portals. In: Proceedings of the 13th International World Wide Web Conference on Alternate track papers & posters. ACM Press, New York (2004)
5. Hyvönen, E., Mäkelä, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., Junnila, M., Kettula, S.: MuseumFinland—Finnish museums on the semantic web. Journal of Web Semantics 3(2), 224–241 (2005)
6. Hyvönen, E., Viljanen, K., Tuominen, J., Seppälä, K.: Building a national semantic web ontology and ontology service infrastructure—the finnonto approach. In: Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M. (eds.) ESWC 2008. LNCS, vol. 5021, pp. 95–109. Springer, Heidelberg (2008)
7. Ding, Y., Fensel, D.: Ontology library systems: The key for successful ontology reuse. In: Cruz, I.F., Decker, S., Euzenat, J., McGuinness, D.L. (eds.) Proceedings of SWWS 2001, The first Semantic Web Working Symposium, Stanford University, California, USA, July 30 - August 1, 2001, pp. 93–112 (2001)
8. Ahmad, M.N., Colomb, R.M.: Managing ontologies: a comparative study of ontology servers. In: ADC 2007: Proceedings of the eighteenth conference on Australasian database, pp. 13–22. Australian Computer Society, Inc., Australia (2007)
9. Viljanen, K., Tuominen, J., Hyvönen, E.: Publishing and using ontologies as mashup services. In: Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M. (eds.) ESWC 2008. LNCS, vol. 5021. Springer, Heidelberg (2008)
10. Euzenat, J., Shvaiko, P.: Ontology matching. Springer, Heidelberg (2007)
11. Mäkelä, E., Viljanen, K., Alm, O., Tuominen, J., Valkeapää, O., Kauppinen, T., Kurki, J., Sinkkilä, R., Känsälä, T., Lindroos, R., Suominen, O., Ruotsalo, T., Hyvänen, E.: Enabling the semantic web with ready-to-use web widgets. In: Proceedings of the First Industrial Results of Semantic Technologies Workshop, ISWC 2007, November 11 (2007)
12. Dameron, O., Noy, N.F., Knublauch, H., Musen, M.A.: Accessing and manipulating ontologies using web services. In: McIlraith, S.A., Plexousakis, D., van Harmelen, F. (eds.) ISWC 2004. LNCS, vol. 3298. Springer, Heidelberg (2004)
13. Vehviläinen, A., Hyvänen, E., 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 (2008)
14. Viljanen, K., Tuominen, J., Känsälä, T., Hyvönen, E.: Distributed semantic content creation and publication for cultural heritage legacy systems. In: Proceedings of the 2008 IEEE International Conference on Distibuted Human-Machine Systems, Athens, Greece. IEEE Press, Los Alamitos (2008)
15. Valkeapää, O., Hyvönen, E., Alm, O.: A framework for ontology-based adaptable content creation on the semantic web. J. of Universal Computer Science (2007)
16. Tuominen, J., Frosterus, M., Viljanen, K., Hyvönen, E.: ONKI SKOS server for publishing and utilizing skos vocabularies and ontologies as services. In: Aroyo, L., et al. (eds.) ESWC 2009. LNCS, vol. 5554, pp. 768–780. Springer, Heidelberg (2009)

17. Kauppinen, T., Henriksson, R., Sinkkilä, R., Lindroos, R., Väätäinen, J., Hyvönen, E.: Ontology-based disambiguation of spatiotemporal locations. In: Proceedings of the 1st international workshop on Identity and Reference on the Semantic Web (IRSW2008), Tenerife, Spain, June 1-5 (2008)
18. Kurki, J.: Finding people and organizations on the semantic web. In: AI and Machine Consciousness - Proceedings of the 13th Finnish Artificial Intelligence Conference STeP, August 20-22 (2008)
19. Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R.S., Peng, Y., Reddivari, P., Doshi, V., Sachs, J.: Swoogle: a search and metadata engine for the semantic web. In: CIKM 2004: Proceedings of the thirteenth ACM international conference on Information and knowledge management, pp. 652–659. ACM Press, New York (2004)
20. d'Aquin, M., Baldassarre, C., Gridinoc, L., Sabou, M., Angeletou, S., Motta, E.: Watson: Supporting next generation semantic web applications. In: Proceedings of IADIS International Conference on WWW/Internet (ICWI 2007), Vila Real, Portugal (2007)
21. Hildebrand, M., van Ossenbruggen, J., Amin, A., Aroyo, L., Wielemaker, J., Hardman, L.: The design space of a configurable autocompletion component. Technical Report INS-E0708, CWI, Amsterdam (November 2007)

Publication II

Jouni Tuominen, Matias Frosterus, Kim Viljanen, and Eero Hyvönen. ONKI SKOS Server for Publishing and Utilizing SKOS Vocabularies and Ontologies as Services. In *The Semantic Web: Research and Applications: 6th European Semantic Web Conference, ESWC 2009, Heraklion, Crete, Greece, May 31–June 4, 2009, Proceedings*, Lora Aroyo, Paolo Traverso, Fabio Ciravegna, Philipp Cimiano, Tom Heath, Eero Hyvönen, Riichiro Mizoguchi, Eyal Oren, Marta Sabou, and Elena Simperl (editors), Lecture Notes in Computer Science, volume 5554, pages 781–795, ISBN 978-3-642-02120-6, Springer-Verlag, June 2009.

ONKI SKOS Server for Publishing and Utilizing SKOS Vocabularies and Ontologies as Services

Jouni Tuominen, Matias Frosterus, Kim Viljanen, and Eero Hyvönen

Semantic Computing Research Group (SeCo)
Helsinki University of Technology (TKK) and University of Helsinki
firstname.lastname@tkk.fi
http://www.seco.tkk.fi/

Abstract. Vocabularies are the building blocks of the Semantic Web providing shared terminological resources for content indexing, information retrieval, data exchange, and content integration. Most semantic web applications in practical use are based on lightweight ontologies and, more recently, on the Simple Knowledge Organization System (SKOS) data model being standardized by W3C. Easy and cost-efficient publication, integration, and utilization methods of vocabulary services are therefore highly important for the proliferation of the Semantic Web. This paper presents the ONKI SKOS Server for these tasks. Using ONKI SKOS, a SKOS vocabulary or a lightweight ontology can be published on the web as ready-to-use services in a matter of minutes. The services include not only a browser for human usage, but also Web Service and AJAX interfaces for concept finding, selecting and transporting resources from the ONKI SKOS Server to connected systems. Code generation services for AJAX and Web Service APIs are provided automatically, too. ONKI SKOS services are also used for semantic query expansion in information retrieval tasks. The idea of publishing ontologies as services is analogous to Google Maps. In our case, however, vocabulary services are provided and mashed-up in applications. ONKI SKOS was published in the beginning of 2008 and is to our knowledge the first generic SKOS server of its kind. The system has been used to publish and utilize some 60 vocabularies and ontologies in the National Finnish Ontology Service ONKI www.yso.fi.

1 Introduction

Thesauri and other controlled vocabularies are used primarily for improving information retrieval [1]. This is accomplished by using concepts or terms of a thesaurus in content indexing, content searching or in both of them, thus simplifying the matching of query terms and the indexed resources (e.g. documents) compared to using natural language. For users, such as content indexers and searchers, to be able to use thesauri, publishing and finding methods for them as well as methods for integrating them with applications [2] are needed [3]. Thesauri are of great benefit for the Semantic Web[1], enabling semantically disambiguated data exchange and integration of data from different sources, though

[1] http://www.w3.org/2001/sw/

L. Aroyo et al. (Eds.): ESWC 2009, LNCS 5554, pp. 768–780, 2009.

not to the same extent as ontologies [4] where the semantics of concepts are defined in more refined and "machine understandable" ways.

Publishing and utilizing thesauri is a laborious task because the representation formats of thesauri and features they provide differ from each other. When utilizing thesauri one has to know how to locate a given thesaurus, and also be familiar with the software the thesaurus is published with. A thesaurus can even be published as a plain text file or even worse, as a paper document, with no proper support for utilizing it. In such a case the users have to implement applications for processing the thesaurus in order to exploit it. Therefore, standard ways for expressing and publishing thesauri would greatly facilitate the publishing and utilizing processes of thesauri.

The Simple Knowledge Organization System (SKOS)[2] [5] being developed within W3C is a data model and a syntax for expressing concept schemes such as thesauri. SKOS provides a standard way for creating vocabularies and migrating existing vocabularies to the Semantic Web. SKOS solves the problem of diverse, non-interoperable thesaurus representation formats by offering a standard convention for presentation. Existing thesauri can be transformed into SKOS format via conversion processes. When a thesaurus is expressed as a SKOS vocabulary, it can be published as a RDF file on the web, allowing the vocabulary users to fetch the files and process them in a uniform way. However, this does not solve the problem of users having to implement their own applications for processing vocabularies.

This paper argues that there are many common, sharable tasks in such vocabulary-aware applications related to e.g. term/concept finding, browsing, selecting, fetching, and query expansion. Lots of work and costs can be saved by implementing such functionalities in standard ways [6] and by providing them for production use as ready-to-use services without having to reimplement the functionalities separately for each local application case. In this way, the use patterns of utilizing vocabularies in the user interface can be harmonized, which makes systems easier to learn and use.

Ontology servers [7,8] have been proposed for publishing ontologies and vocabularies on the Semantic Web. Such servers are used for managing ontologies and offering users access to them. For accessing SKOS vocabularies, there is the Web Service based SKOS API[3] developed in the SWAD-Europe project, and the terminology service by Tudhope et al. [9]. There is also ongoing research by the Networked Knowledge Organization Systems/Services (NKOS) community[4] that focuses on enabling knowledge organization systems as networked interactive information services to support the description and retrieval of diverse information resources through the Internet. However, general tools for providing out-of-the-box support for utilizing SKOS vocabularies in, e.g., content indexing, without needing to implement application specific user interfaces for end users do not exist.

[2] http://www.w3.org/TR/skos-reference/
[3] http://www.w3.org/2001/sw/Europe/reports/thes/skosapi.html
[4] http://nkos.slis.kent.edu/

This paper fills this gap by presenting the ONKI SKOS Server for publishing and utilizing thesauri and lightweight ontologies. In the following, ways of presenting thesauri on the Semantic Web are first overviewed. After this, the approach and implementation of the ONKI SKOS Server is presented followed by descriptions of use cases of its services. In conclusion, contributions of the work are summarized and directions for further research outlined.

2 Presenting Thesauri on the Semantic Web

W3C's SKOS data model provides a vocabulary for expressing the basic structure and contents of concept schemes, such as thesauri, classification schemes and taxonomies. The concept schemes are expressed as RDF graphs by using classes and properties defined in the SKOS specification, thus making thesauri compatible with the Semantic Web. SKOS is capable of representing vocabularies which loosely conform to the influential ISO 2788 thesaurus standard [6].

Although semantically richer RDFS/OWL ontologies enable more extensive ways to perform logical inferencing than SKOS vocabularies, in several cases thesauri represented with SKOS are sufficient. In our opinion, the first and the most obvious benefit of using Semantic Web ontologies/vocabularies in content indexing is their ability to disambiguate concept references in a universal way. This is achieved by using persistent URIs as the identification mechanism and is a tremendous advantage when compared with the traditional idea of controlled vocabularies where plain concept labels are used as identifiers. In such a case, identification problems can be encountered e.g. with polysemous and homonymous labels, when dealing with multi-lingual resources, and with resources whose labels may change as time goes by. For example, human names are transliterated in many ways in different languages, may change due to marriage, a person may have nick names, and so on. In such cases the labels of concepts are not a permanent identification method, and the references to the concepts may become ambiguous or invalid.

URIs provide not only an identification mechanism, but also means for accessing the concept definitions and thesauri, when using the http URI Scheme[5]. Such URIs can act as URLs, and with a proper server configuration provide the users additional information about the referred concept[6]. In addition to these general RDF characteristics, SKOS provides a way for expressing relations between concepts suitable for the needs of thesauri, thus providing conceptual context for concepts.

In short, using a common representation model such as SKOS for thesauri greatly reduces the cost of (a) sharing thesauri, (b) using different thesauri in conjunction within one application, and (c) development of standard software to process them [6].

[5] `http://www.iana.org/assignments/uri-schemes.html`
[6] `http://www.w3.org/TR/swbp-vocab-pub/`

3 Utilizing Thesauri with ONKI SKOS Services

ONKI SKOS is an ontology server implementation for publishing and utilizing thesauri and lightweight ontologies. It conforms to the general ONKI vision and API, and is thus usable via ONKI ontology services as easily integrable user interface components and APIs [2].

The Semantic Web applications typically use ontologies which are either straightforward conversions of well-established thesauri, application-specific vocabularies or semantically richer ontologies, that can be presented and accessed in similar ways to thesauri [3,10]. Since SKOS defines a suitable model for expressing thesauri, it was chosen as the primary data model supported by the ONKI SKOS Server.

ONKI SKOS can be used to browse, search and visualize any vocabulary conforming to the SKOS specification, and also RDFS/OWL ontologies with additional configuration. ONKI SKOS does simple reasoning, e.g., transitive closure over class and part-of hierarchies. The implementation has been piloted using various thesauri and ontologies from diverse domains. Piloted ontologies include ontologies for cultural heritage (Art & Architecture Thesaurus AAT[7], 28,000 concepts; Iconclass[8], 27,000 concepts), health ontologies (Medical Subject Headings MeSH[9], 24,000 concepts), business ontologies (United Nations Standard Products and Services Code UNSPSC[10], 21,000 concepts), geographical instance ontologies (The Getty Thesaurus of Geographic Names TGN[11], 143,000 concepts), upper ontologies (General Finnish Upper Ontology YSO [3], 21,000 concepts), governmental ontologies (Finnish Governmental Thesaurus VNAS, 6,300 concepts) and natural science taxonomies (Ontology of Birds of the World AVIO, 11,000 concepts; Ontology of Mammals of the World MAMO, 5,000 concepts). However, not all of these ontologies are freely accessible due to licensing issues.

When utilizing thesauri represented as SKOS vocabularies and published on the ONKI SKOS server, several benefits are gained. Firstly, SKOS provides a universal way of expressing thesauri. Thus processing different thesauri can be done in the same way, eliminating the use of thesaurus specific processing rules in applications or separate converters between various formats. Secondly, ONKI SKOS provides access to all published thesauri in the same way, so content indexers and end-users do not have to use thesaurus specific implementations of thesaurus browsers and other tools developed by different parties, which is the predominant way. Also, one of the goals of the ONKI ontology services is that all the essential ontologies/thesauri can be found at the same location, thus eliminating the need to search for other thesaurus sources.

The typical way to use thesaurus specific publishing systems in content indexing and searching is either by using their browser user interface for finding

[7] http://www.getty.edu/research/conducting_research/vocabularies/aat/
[8] http://www.iconclass.nl/
[9] http://www.nlm.nih.gov/mesh/meshhome.html
[10] http://www.unspsc.org/
[11] http://www.getty.edu/research/conducting_research/vocabularies/tgn/

desired concepts and then copying and pasting the concept label to the used indexing system[12], or by using APIs for accessing and querying the thesaurus [9]. Both methods have some drawbacks. The first method introduces rather uncomfortable task of constant switching between two applications and the clumsy copy-paste procedure. The second method leaves the implementation job of the user interface entirely to the parties utilizing the thesaurus.

While ONKI SKOS supports both the aforementioned thesauri utilizing methods, in addition, as part of the ONKI ontology services, it provides a lightweight web widget for integrating general thesauri accessing functionalities into web based applications on the user interface level. The ONKI Selector widget depicted in Figure 1 can be used to search and browse thesauri, fetch URI references and labels of desired concepts and storing them in a concept collector. Similar ideas have been proposed by Hildebrand et al. [11] for providing search widget for general RDF repositories, and by Vizine-Goetz et al. [12] for providing widget for accessing thesauri through the side bar of the Internet Explorer web browser.

When the desired concepts have been selected with the ONKI Selector they can be stored into, e.g., the database of the application by using an HTML form. Either the URIs or the labels of the concepts can be transferred into the application providing support for the Semantic Web and legacy applications. For browsing the context of the concepts in thesauri, the ONKI SKOS Browser can be opened by pressing a button. Desired concepts can be fetched from the browser to the application by pressing the "Fetch Concept" button. Thus, there is no need for copy-paste procedures or user interface implementation projects.

For information retrieval use cases, ONKI SKOS provides support for finding suitable query terms and expanding them by using ontological inference, e.g., based on the concept subsumption, partonymy or associative relations between concepts. In legacy systems, which do not support URIs, the labels of the concepts can be used as query terms. In this context, ONKI SKOS supports multilingual search. If a thesaurus contains, e.g., English and Finnish labels for the concepts, the relevant query concepts can be searched in English, while in the actual information retrieval task the used query terms are in Finnish. This is useful when the end-user does not understand the language of the terms used in indexing the content of the underlying system.

The ONKI SKOS Browser (see Figure 2) is the graphical user interface of ONKI SKOS. It consists of three main components: 1) *semantic autocompletion concept search*, 2) *concept hierarchy* and 3) *concept properties*. When typing text to the search field, a query is performed to match the concepts' labels. The result list shows the matching concepts, which can be selected for further examination.

When a concept is selected, its concept hierarchy is visualized as a tree structure. The ONKI SKOS Browser supports multi-inheritance of the concepts (i.e. a concept can have multiple parents). Whenever a multi-inheritance structure is met, a new branch is formed to the tree. This leads to cloning of nodes, i.e.

[12] This is the way the Finnish General Thesaurus YSA has been used previously via the VESA Web Thesaurus Service, `http://vesa.lib.helsinki.fi/`

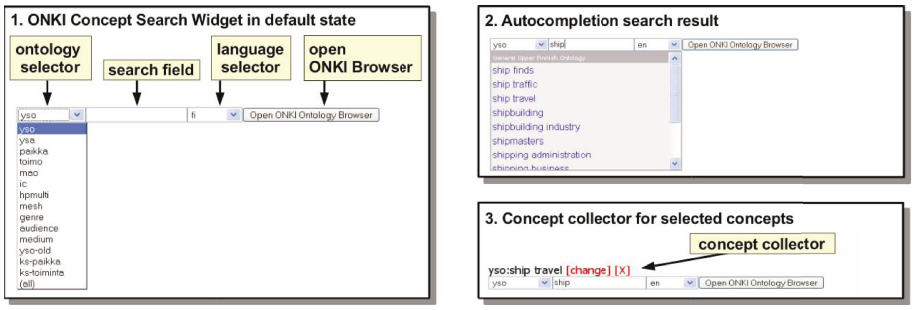


Fig. 1. ONKI Selector for concept finding

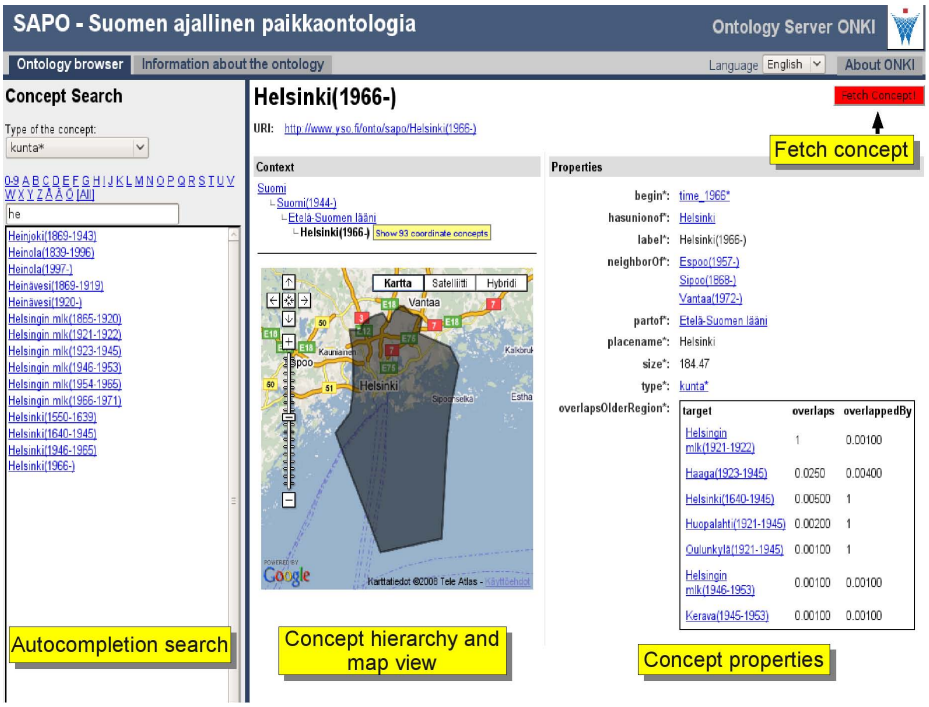


Fig. 2. The ONKI SKOS Browser

a concept can appear multiple times in the hierarchy tree. Next to the concept hierarchy tree, the properties of the selected concept are shown in the user interface. Mappings and other relations between concepts act as links allowing user to browse thesauri. In context of geographical ontologies, the place instances can be visualized on a Google Maps[13] view, as depicted in the Figure 2, if the instances contain coordinate information.

[13] `http://maps.google.com/`

The National Finnish Ontology Service ONKI[14] provides documentation and tools for software architects developing ontology-based applications and integrating ONKI services into them [13]. With ONKI Selector Builder a developer can set the desired configuration properties for the ONKI Selector and generate the needed JavaScript code to integrate the ONKI Selector into an application. Also, a hands-on demonstration page for integrating the ONKI Selector is available. The page contains a HTML form, into which the user can type HTML/JavaScript code and see how the resulting page will look like in the end-user's web browser.

The Web Service and AJAX interfaces of the ONKI SKOS server can be used for querying for concepts by label matching, performing query expansion, getting label for a given URI and for querying for supported languages and type URIs of a thesaurus. For testing Web Service methods, a demonstration page is provided at the ONKI service. One can try out the methods of the ONKI API and examine the resulting SOAP request and response messages.

ONKI SKOS is implemented as a Java Servlet using the Jena Semantic Web Framework[15], the Direct Web Remoting library[16] and the Lucene[17] text search engine.

4 Configuring ONKI SKOS with SKOS Structures

ONKI SKOS supports straightforward loading of SKOS vocabularies with minimal configuration needs. For using other data models than SKOS, various configuration properties are specified to enable ONKI SKOS to process the thesauri/ontologies as desired. The configurable properties include the properties used in hierarchy generation, the properties used to label the concepts, the concept to be shown in the default view and the default concept type used in restricting the concept search.

When the ONKI SKOS Browser is accessed with no URL parameters, information related to the concept configured to be shown as default is shown. Usually this resource is the root resource of the vocabulary, if the vocabulary forms a full-blown tree hierarchy with one single root. In SKOS concept schemes the root resource is the resource representing the concept scheme itself, i.e. the instance of *skos:ConceptScheme*.

The concept hierarchy of a concept is generated by traversing the configured properties. In SKOS these properties are *skos:narrower* and *skos:broader* and they are used to express the hierarchical relations between concepts. Hierarchical relations between the root resource representing the concept scheme and the top concepts of the concept scheme are defined with the property *skos:hasTopConcept*.

Labels of concepts are needed in visualizing search results, concept hierarchies, and related concepts in the concept property view. In SKOS the labels are

[14] `http://www.yso.fi/`
[15] `http://jena.sourceforge.net/`
[16] `http://directwebremoting.org/`
[17] `http://lucene.apache.org/java`

expressed with the property *skos:prefLabel*. The label is of the same language as the currently selected user interface language, if such a label exists. Otherwise any label is used.

The semantic autocompletion search of ONKI SKOS works by searching for concepts whose labels match the search string. To support this, the labels of the concepts are indexed. The indexed properties can be configured. In SKOS these properties are *skos:prefLabel*, *skos:altLabel* and *skos:hiddenLabel*. When the user searches, e.g., with the search term "cat", all concepts which have one of the aforementioned properties with values starting with the string "cat" are shown in the search results. The autocompletion search also supports wildcards, so a search with a string "*cat" returns the concepts which have the string "cat" as any part of their label.

If a SKOS vocabulary contains concepts that are instances of subclasses of skos:Concept, the search can be limited to certain types of concepts only. To accomplish this, the types of the concepts (which are expressed with the property *rdf:type*) are indexed. It is also possible to limit the search to a certain subtree of the concept hierarchy by restricting the search to subconcepts of a specific concept. To support this, also the parents of concepts are indexed.

Many thesauri include structures for representing categories of concepts. To support category-based concept search in the ONKI SKOS Browser, another search field is provided. When a category is selected from the category search view, the concept search is restricted to the concepts belonging to the selected category. SKOS includes a concept collection structure, *skos:Collection*, which can be used for expressing such categories. However, *skos:Collection* is often used for slightly different purposes, namely for node labels[18]. Thus, instances of *skos:Collection* are not used for category-based concept search by default.

5 Converting Thesauri to SKOS—Case YSA

Publishing a thesaurus in the ONKI SKOS server is straightforward. To load a SKOS vocabulary into the server, only the location path of the RDF file of the vocabulary needs to be configured manually. After rebooting the ONKI SKOS, the RDF file is loaded, indexed and made accessible for users. ONKI SKOS provides the developers of thesauri a simple way to publish their thesauri.

There exists quite amount of well-established keyword lists, thesauri and other non-RDF controlled vocabularies which have been used in traditional approaches in harmonizing content indexing. In order to reuse the effort already invested developing these resources by publishing these vocabularies in ONKI SKOS server, conversion processes need to be developed. This idea has also been suggested by van Assem et al. [6] and Summers et al. [14]. We have implemented transformation scripts for, e.g., MARCXML format[19], XML dumps from SQL databases and proprietary XML schemas.

[18] A construct for displaying grouping concepts in systematic displays of thesauri. They are not actual concepts, and thus they should not be used for indexing. An example node label is "milk by source animal".

[19] `http://www.loc.gov/standards/marcxml/`

An example of the SKOS transformation and publishing process is the case of YSA, the Finnish General Thesaurus[20]. YSA is developed by the National Library of Finland and is the most widely used controlled vocabulary in Finland. YSA is published in the VESA Web Thesaurus Service, where it receives over 12 million user hits per year. YSA is simultaneously published in the ONKI SKOS, currently in a pilot phase. During this year, the operation of the VESA service will be ended, and the ONKI SKOS becomes the official publishing channel of YSA.

To begin the SKOS transformation process, YSA is exported into MARCXML format from the thesaurus management system of the National Library of Finland. The resulting constantly up-to-date file is published at the web server of the National Library of Finland, from where it is fetched via OAI-PMH protocol[21] to our server. This process is automated and the new version of the XML file is fetched daily. Instead of fetching the full file every day, the incremental update feature of the OAI-PMH could be used, but we have not tested it. After fetching a new version of the file, the transformation process depicted in Figure 3 is started by loading the MARCXML file (ysa.xml). The Java-based converter first creates the necessary structure and namespaces for the SKOS model utilizing Jena Semantic Web Framework. Next, the relations in YSA are translated into their respective SKOS counterparts, which is depicted in Figure 4.

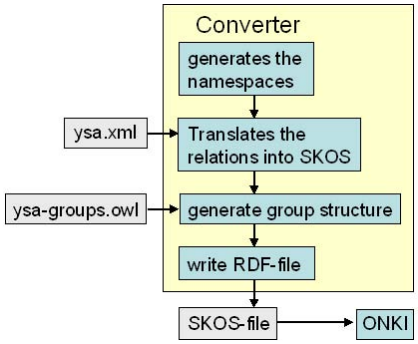


Fig. 3. The SKOS transformation process of YSA

A URI for the new concept entry is created through the unique ID in the source file. The preferred and alternative labels are converted straightforwardly from one syntax to another. Similarly the type and scheme definitions are added to the SKOS model. Since the relations in the MARCXML refer not to the identifiers but rather to the labels, the source file is searched for an entry that has the given label and then its ID is recorded for the SKOS relation.

MARCXML record for the concept "pistols" can be seen in the left part of Figure 4. The ID of the concept is denoted by the *controlfield* with *tag* attribute's

[20] `http://www.nationallibrary.fi/libraries/thesauri/ysa.html`
[21] `http://www.openarchives.org/OAI/openarchivesprotocol.html`

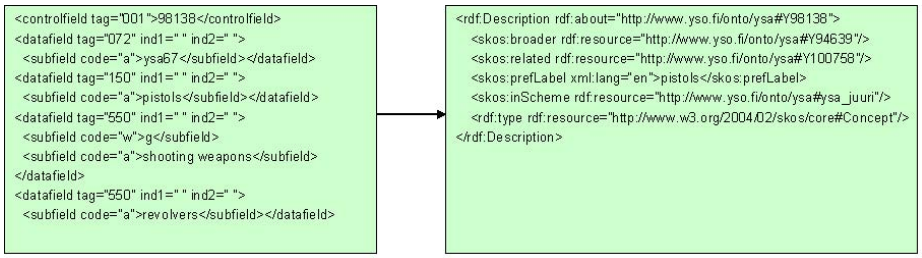


Fig. 4. An example of the SKOS transformation of YSA

value "001" ("98138"). This ID value becomes the local name of the URI of the corresponding SKOS concept. The category in which the concept belongs to is depicted with the *datafield* with *tag* "072" ("ysa67"). The preferred label is the value provided in the *datafield* with *tag* value "150" ("pistols"). Hierarchical and associative relations are marked with *datafield* with *tag* value "550", while the *subfield*'s attribute *code* specifies which one of the relations is in question.

Once the SKOS transformation is ready, the converter fetches the labels for the concept categories from a separate file (ysa-groups.owl) - these labels are not included in the MARCXML file. Finally, a RDF file is written and imported into ONKI SKOS.

6 Use Cases of ONKI SKOS

ONKI SKOS has been tested in several content indexing and information retrieval pilot applications.

6.1 Content Indexing

The browser-based annotation editor SAHA[22] [15] has been used for creating semantic content for semantic portals. The two major applications demonstrating Semantic Web technologies are HealthFinland[23] [16] in the health promotion context, and CultureSampo[24] [17] in the cultural heritage context. Both systems use the SAHA editor with the services provided by ONKI SKOS and the National Finnish Ontology Service ONKI. ONKI Selector has been integrated into SAHA for finding suitable concepts in content indexing.

The web laboratory Owela[25] of the VTT Technical Research Centre of Finland has implemented a service for collecting and sharing text and image clips from the Web[26]. In the service one can organize the clips into folders and tag them with

[22] http://demo.seco.tkk.fi/saha/?lang=en
[23] http://demo.seco.tkk.fi/tervesuomi/
[24] http://www.kulttuurisampo.fi/
[25] http://owela.vtt.fi/
[26] http://owela.vtt.fi/tilkut

different categories. The ONKI Selector is used for tagging the clips keywords from vocabularies published in the ONKI service.

6.2 Information Retrieval

Kantapuu.fi[27] is a web user interface for browsing and searching for collections of Finnish museums of forestry. The collection items are annotated with Finnish terms from General Finnish Thesaurus YSA, Thesaurus for Museum Domain MASA and Agriforest Thesaurus. Kantapuu.fi search page is a web form into which query strings are typed as free text. The query strings can be placed into specific fields, e.g. "keywords", "place of use" or "time of use". We have created a demonstration page containing a Kantapuu.fi's search form with integrated ONKI Selectors which can be used for selecting query terms to be used in the Kantapuu.fi search[28]. The ONKI Selector is used for finding terms from vocabularies of the ONKI service. The used vocabularies are the same as those used in the annotation process of the items, or actually their ontologized versions. To find suitable query terms user can utilize the autocompletion search functionality or the ONKI Browser. Thus, the user does not need to be familiar with the vocabularies used in the annotations of the items, as in the case of free text search. The ONKI Selector performs query expansion based on the selected query terms. So, for example a query term "animals" would return items annotated with term "cats". When the desired query terms are selected, the actual search to the Kantapuu.fi system can be executed. The users of the pilot system have given positive feedback especially on the multilingual search possibility.

The ONKI Selector Widget has also been integrated into the Viikki Science Library[29] reference database system eViikki[30]. eViikki is a search interface for the library's collections, which consist of scientific literature on agriforestry. The literature is annotated with terms from General Finnish Thesaurus YSA and Agriforest vocabulary. The ONKI Selector is used for populating the "keywords" field of the search form of eViikki. The fetched concept labels are used in the information retrieval task. Query expansion is not performed currently.

7 Discussion

The main contribution of this paper is to show, both in principle and as a deployed implemented service in use, how thesauri can be published and utilized easily on the Semantic Web, emphasizing the benefits of the use of W3C's SKOS data model as a uniform vocabulary representation framework. The ONKI SKOS server is presented as a proof of concept for a cost-efficient thesauri utilization method. By using ONKI SKOS, general thesauri accessing functionalities can be easily integrated into applications without the need for users to reimplement their own user interfaces for this.

[27] http://www.kantapuu.fi/
[28] http://www.yso.fi/lusto
[29] http://www.tiedekirjasto.helsinki.fi/english/
[30] http://www-db.helsinki.fi/eviikki/eviikkihaku.html

The utilization of the SKOS structures in an ontology server was described in context of the ONKI SKOS server. The case of the Finnish General Thesaurus acts as one example on how an existing, widely used thesaurus can be converted into the SKOS format and be published on the ONKI SKOS server. At the moment some 60 international and national vocabularies, taxonomies, and lightweight ontologies have been published online using ONKI SKOS and the number is increasing. Finally, in order to demonstrate end-user benefits of ONKI SKOS services, we provided some real world use cases of ONKI SKOS in content indexing and information retrieval.

Future work for ONKI SKOS includes creating a more extensive Web Service interface for supporting, e.g., querying for properties of a given concept and for discovering concepts which are related to a given concept. The starting point for this API will be the SKOS API. Also, version control techniques and other support for vocabulary management (e.g. protocol for communicating vocabulary changes to users) would be needed to support the development phase of vocabularies.

Acknowledgements

We thank Ville Komulainen for his work on the original ONKI server. This work is a part of the National Semantic Web Ontology project in Finland[31] (FinnONTO) and its follow-up project Semantic Web 2.0[32] (FinnONTO 2.0, 2008-2010), funded mainly by the National Technology and Innovation Agency (Tekes) and a consortium of 38 private, public and non-governmental organizations.

References

1. Aitchison, J., Gilchrist, A., Bawden, D.: Thesaurus Construction and Use: A Practical Manual, 4th edn. Europa Publications (2000)
2. Viljanen, K., Tuominen, J., Hyvönen, E.: Publishing and using ontologies as mashup services. In: Proceedings of the 4th Workshop on Scripting for the Semantic Web (SFSW 2008), Tenerife, Spain, June 1-5 (2008)
3. Hyvönen, E., Viljanen, K., Tuominen, J., Seppälä, K.: Building a national semantic web ontology and ontology service infrastructure—the FinnONTO approach. In: Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M. (eds.) ESWC 2008. LNCS, vol. 5021, pp. 95–109. Springer, Heidelberg (2008)
4. Staab, S., Studer, R. (eds.): Handbook on ontologies. Springer, Heidelberg (2004)
5. Miles, A., Matthews, B., Wilson, M., Brickley, D.: SKOS Core: Simple knowledge organisation for the web. In: Proceedings of the International Conference on Dublin Core and Metadata Applications (DC 2005), Madrid, Spain, September 12-15 (2005)

[31] `http://www.seco.tkk.fi/projects/finnonto/`
[32] `http://www.seco.tkk.fi/projects/sw20/`

6. van Assem, M., Malaisé, V., Miles, A., Schreiber, G.: A method to convert thesauri to SKOS. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, pp. 95–109. Springer, Heidelberg (2006)
7. Ding, Y., Fensel, D.: Ontology library systems: The key to successful ontology reuse. In: Proceedings of SWWS 2001, The first Semantic Web Working Symposium, Stanford University, USA, pp. 93–112 (August 2001)
8. Ahmad, M.N., Colomb, R.M.: Managing ontologies: a comparative study of ontology servers. In: Proceedings of the eighteenth Conference on Australasian Database (ADC 2007), Ballarat, Victoria, Australia, January 30–February 2, pp. 13–22. Australian Computer Society, Inc (2007)
9. Tudhope, D., Binding, C.: Towards terminology service: experiences with a pilot web service thesaurus browser. In: Proceedings of the International Conference on Dublin Core and Metadata Applications (DC 2005), Madrid, Spain, September 12-15, 2005, pp. 269–273 (2005)
10. van Assem, M., Menken, M.R., Schreiber, G., Wielemaker, J., Wielinga, B.: A method for converting thesauri to RDF/OWL. In: McIlraith, S.A., Plexousakis, D., van Harmelen, F. (eds.) ISWC 2004. LNCS, vol. 3298, pp. 17–31. Springer, Heidelberg (2004)
11. Hildebrand, M., van Ossenbruggen, J., Amin, A., Aroyo, L., Wielemaker, J., Hardman, L.: The design space of a configurable autocompletion component. Technical Report INS-E0708, Centrum voor Wiskunde en Informatica (CWI), Amsterdam (2007)
12. Vizine-Goetz, D., Childress, E., Houghton, A.: Web services for genre vocabularies. In: Proceedings of the International Conference on Dublin Core and Metadata Applications (DC 2005), Madrid, Spain, September 12-15 (2005)
13. Viljanen, K., Tuominen, J., Hyvönen, E.: Ontology libraries for production use: The Finnish ontology library service ONKI. In: Aroyo, L., et al. (eds.) ESWC 2009. LNCS, vol. 5554, pp. 781–795. Springer, Heidelberg (2009)
14. Summers, E., Isaac, A., Redding, C., Krec, D.: LCSH, SKOS and Linked Data. In: Proceedings of the International Conference on Dublin Core and Metadata Applications (DC 2008), Berlin, Germany, September 22-26 (2008)
15. Valkeapää, O., Alm, O., Hyvönen, E.: Efficient content creation on the semantic web using metadata schemas with domain ontology services (system description). In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, pp. 819–828. Springer, Heidelberg (2007)
16. Hyvönen, E., Viljanen, K., Suominen, O.: HealthFinland—Finnish health information on the semantic web. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 778–791. Springer, Heidelberg (2007)
17. Hyvönen, E., Ruotsalo, T., Häggström, T., Salminen, M., Junnila, M., Virkkilä, M., Haaramo, M., Mäkelä, E., Kauppinen, T., Viljanen, K.: Culturesampo—Finnish culture on the semantic web: The vision and first results. In: Robering, K. (ed.) Information Technology for the Virtual Museum. LIT Verlag, Berlin (2007)

Publication III

Jouni Tuominen, Tomi Kauppinen, Kim Viljanen, and Eero Hyvönen. Ontology-Based Query Expansion Widget for Information Retrieval. In *Proceedings of the 5th International Workshop on Scripting and Development for the Semantic Web at ESWC 2009*, Heraklion, Greece, May 31, Sören Auer, Chris Bizer, and Gunnar Aastrand Grimnes (editors), CEUR Workshop Proceedings, volume 449, pages 52–57, ISSN 1613-0073, online CEUR-WS.org/Vol-449/ShortPaper1.pdf, May 2009.

Ontology-Based Query Expansion Widget for Information Retrieval

Jouni Tuominen, Tomi Kauppinen, Kim Viljanen, and Eero Hyvönen

Semantic Computing Research Group (SeCo)
Helsinki University of Technology (TKK) and University of Helsinki
http://www.seco.tkk.fi/
firstname.lastname@tkk.fi

Abstract. In this paper we present an ontology-based query expansion widget which utilizes the ontologies published in the ONKI Ontology Service. The widget can be integrated into a web page, e.g. a search system of a museum catalogue, enhancing the page by providing a query expansion functionality. We have tested the system with general, domain-specific and spatio-temporal ontologies.

1 Introduction

In information retrieval systems the relevancy of search results depends on the user's ability to represent her information needs in a query [1]. If the vocabularies used by the user and the system are not the same ones, or if the shared vocabulary is used in different levels of specificity, the search results are usually poor. Query expansion has been proposed to solve these issues and to improve information retrieval by expanding the query with terms related to the original query terms. Query expansion can be based on corpus, e.g. analyzing co-occurences of terms, or on knowledge models, such as thesauri [2] or ontologies [1]. Methods based on knowledge models are especially useful in cases of short, incomplete query expressions with few terms found in the search index [1, 2].

We have implemented a web widget providing query expansion functionality to web-based systems as an easily integrable service with no need to change the underlying system. The widget uses ontologies to expand the query terms with semantically related concepts. The widget extends the previously developed ONKI Selector widget, which is used for selecting concepts especially for annotation purposes [3].

The user does not have to be familiar with the ontologies used in content annotations by utilizing the autocompletion search feature of the widget, as the system suggests matching concepts as the user is writing the query string. Also, to help the user to disambiguate concepts the ONKI Ontology Browsers [4] can be used to get a better understanding of the semantics of the concepts, e.g. by providing a concept hierarchy visualization.

The query expansion widget supports Semantic web and legacy systems[1], i.e. either the concept URIs or the concept labels can be used in queries. In

[1] By legacy systems we mean systems that do not use URIs as identifiers.

legacy systems cross-language search can be performed, if the used ontology contains concept labels in several languages. In addition to the widget, the query expansion service can also be utilized via JavaScript and Web Service APIs. The query expansion widget and the APIs are available for public use as part of the ONKI Ontology Service[2] [4]. The JavaScript code needed for integrating the widget into a search system can be generated by using the ONKI Widget Generator[3].

The contribution of this paper is to present an approach to perform query expansion in systems cost-effectively, not to evaluate how the chosen query expansion methods improve information retrieval in the systems.

2 Ontologies used for Query Expansion

The ONKI query expansion widget can be used with any ontology published in the ONKI Ontology Service. The service contains some 60 ontologies at the time of writing. Users are encouraged to submit their own ontologies to be published in the service by using the Your ONKI Service[4]. In the following, we describe how we have used different types of ontologies for query expansion.

2.1 Query Expansion with General and Domain-specific Ontologies

For expanding general and domain-specific concepts in queries we have used The Finnish Collaborative Holistic Ontology KOKO[5] which consists of The Finnish General Upper Ontology YSO [5] and several domain-specific ontologies expanding it. To improve poor search results caused by using vocabularies in different levels of specificity in queries and in the search index we have used the transitive is-a relation (*rdfs:subClassOf*[6]) for expanding the query concepts with their subclasses. So for example, when selecting a query concept *publications*, the query is expanded with concepts *magazines*, *books*, *reports* and so on.

Using other relations in addition or instead of the is-a relation in query expansion might be beneficial. When considering general associative relations, caution should be exercised as their use in query expansion can lead to uncontrolled expansion of result sets, and thus to potential loss in precision [6, 7]. In case of a legacy system (not handling URIs, using labels instead) the use of alternative labels of concepts (synonyms) may improve the search. The relations used in the query expansion of an ontology can be configured when publishing the ontology in the ONKI Ontology Service.

[2] `http://www.yso.fi/`
[3] `http://www.yso.fi/onkiselector/`
[4] `http://www.yso.fi/upload/`
[5] `http://www.seco.tkk.fi/ontologies/koko/`
[6] Defined in the RDFS Recommendation, `http://www.w3.org/TR/rdf-schema/`

2.2 Query Expansion with the Spatio-temporal Ontology SAPO

A spatial query can explicitly contain spatial terms (e.g. Helsinki) and spatial relations (e.g. near), but implicitly it can include even more spatial terms that could be used in query expansion [8]. For example, in a query "museums near Helsinki" not only Helsinki is a relevant spatial term, but also its neighboring municipalities. Spatial terms – i.e. geographical places – do not exist just in space but also in time [9, 10]. This is especially true for museum collections where objects have references to places from different times. This sets a requirement to utilize also relations between historical places and more contemporary places in query expansion. To provide these mappings we used a spatio-temporal ontology SAPO (The Finnish Spatio-temporal Ontology) [11].

In SAPO regional overlap mappings are expressed as depicted in Figure 1, where example Turtle RDF[7] statements[8] express that the region of the latest temporal part of place *sapo:Joensuu* — i.e. the one valid from the beginning of year 2009 — overlaps the region of the temporal part of *sapo:Eno* of years 1871–2008. The temporal part of the place simply means the place during a certain time-period such that different temporal parts might have different extensions (i.e. borders) [11].

```
sapo:Joensuu(2009-)
    sapo:begin
            "2009-01-01" ;
    sapo:overlaps
            sapo:Eno(1871-2008) ,
            sapo:Pyhaselka(1925-2008) ,
            sapo:Joensuu(2005-2008) .
```

Fig. 1. Overlap mappings between temporal parts of places.

```
sapo:Joensuu
     sapo:unionof
             sapo:Joensuu(1848-1953) ,
             sapo:Joensuu(1954-2004) ,
             sapo:Joensuu(2005-2008) ,
             sapo:Joensuu(2009-) ;
     sapo:overlapsAtSomeTime
             sapo:Eno ,
             sapo:Pyhaselka ,
             sapo:Tuupovaara ,
             sapo:Pielisensuu ,
             sapo:Kiihtelysvaara .
```

Fig. 2. A place is a union of its temporal parts. Moreover, places may have overlapped other places *at some time*.

For example, the place *sapo:Joensuu* is a union of four temporal parts, defined in the example depicted in Figure 2. However, annotations of items likely utilize places rather than their temporal parts. For this reason the model uses property *sapo:overlapsAtSomeTime* to explicate that e.g. a place *sapo:Joensuu* has — at some point in the history — overlapped together five different places (*sapo:Eno* and four others). In other words, e.g. at least one temporal part of *sapo:Joensuu* has overlapped at least one temporal part of *sapo:Eno*. We have used this more generic property *sapo:overlapsAtSomeTime* between places for query expansion.

[7] `http://www.dajobe.org/2004/01/turtle/`

[8] The example uses the following prefix - sapo: http://www.yso.fi/onto/sapo/

3 A Use Case of the Query Expansion Widget

We have created a demonstration search interface[9] consisting of the original Kantapuu.fi search form[10] and integrated ONKI widgets for query expansion. Kantapuu.fi is a web user interface for browsing and searching for collections of Finnish museums of forestry, using simple matching algorithm of free text query terms with the item index terms. The ontologies used in the query expansion are the same ones as used in annotation of the items[11], namely The Finnish General Upper Ontology YSO, Ontology for Museum Domain MAO[12] and Agfiforest Ontology AFO[13]. For expanding geographical places the Finnish Spatio-temporal Ontology SAPO is used.

When a desired query concept is selected from the results of the autocompletion search of the widget or by using the ONKI Ontology Browser, the concept is expanded. The resulting query expression is the disjunction of the original query concept and the concepts expanding it, formed using the Boolean operation OR. The query expression is placed into a hidden input field, which is sent to the original Kantapuu.fi search page when the HTML form is submitted.

An example query is depicted in Figure 3, where the user is interested in old publications from place Joensuu. User has used the autocompletion feature of the widget to input to the *keywords* field a query term "publicat", which has been autocompleted to the concept *publications*, which has been further expanded to its subclasses (their Finnish labels). Similarly, the place *Joensuu* has been added to the field *place of usage* and expanded with the places it overlaps.

The result set of the search contains four items, from which two are magazines used in place Eno and the rest two are cabinets for books used in place Joensuu. Without using the query expansion the result set would have been empty, as the place *Eno* and the concept *books* were not in the original query.

4 Discussion

When implementing the demonstration search interface for the Kantapuu.fi system with ONKI widgets we faced some challenges. If a query concept has lots of subconcepts, the expanded query string may become inconveniently long, as the concept URIs/labels of the subconcepts are added to the query. This may cause problems because the used HTTP server, database system or other software components may set limits to the length of the query string. With lengthy queries the system may not function properly or the response times of the system may increase.

[9] `http://www.yso.fi/kantapuu-qe/`
[10] `http://www.kantapuu.fi/`, follow the navigation link "Kuvahaku".
[11] To be precise, the ontologies are based on thesauri that have been used in annotation of the items.
[12] `http://www.seco.tkk.fi/ontologies/mao/`
[13] `http://www.seco.tkk.fi/ontologies/afo/`

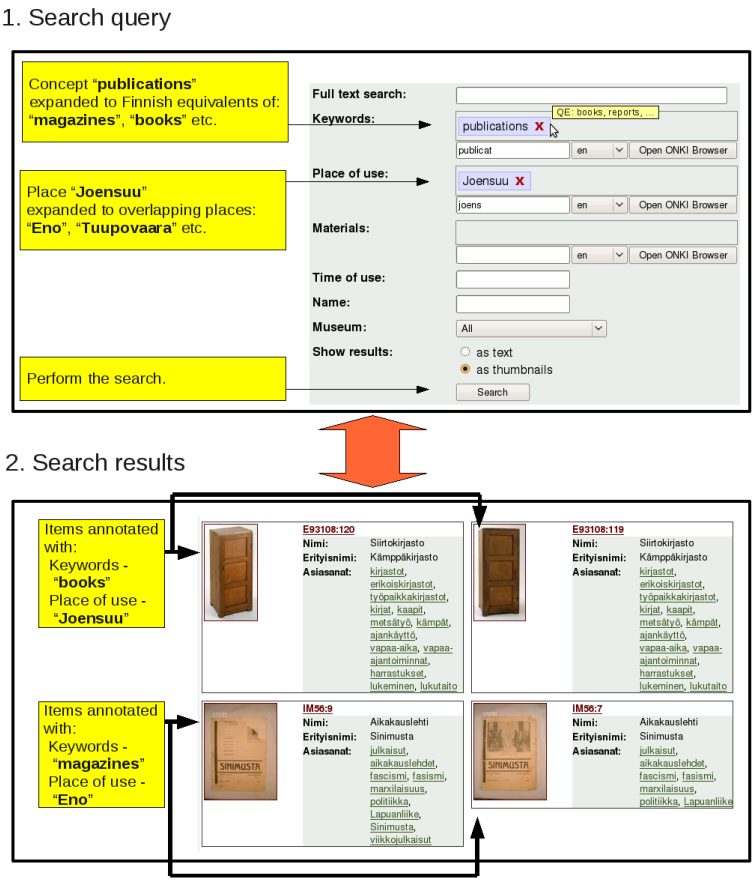


Fig. 3. Kantapuu.fi system with integrated ONKI widgets.

Future work includes user testing for finding out if users consider the query expansion of the concepts and places useful. Also, systematic evaluation of the search systems used would be essential to find out if the query expansion improves the information retrieval, and specifically which semantic relations improve the results the most. The user interface of the query expansion widget needs further developing, e.g., the user should be able to select/unselect the suggested query expansion concepts.

Acknowledgements

We thank Ville Komulainen for his work on the original ONKI server and Leena Paaskoski and Leila Issakainen for cooperation on integrating the ONKI query expansion widgets into the Kantapuu.fi system. This work has been partially

funded by Lusto The Finnish Forest Museum[14] and partially by the IST funded EU project SMARTMUSEUM[15] (FP7-216923). The work is a part of the National Semantic Web Ontology project in Finland[16] (FinnONTO) and its follow-up project Semantic Web 2.0[17] (FinnONTO 2.0, 2008-2010), funded mainly by the National Technology and Innovation Agency (Tekes) and a consortium of 38 private, public and non-governmental organizations.

References

1. 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, Dublin, Ireland (July 3-6 1994) 61–69
2. Wang, Y.C., Vandendorpe, J., Evens, M.: Relational thesauri in information retrieval. Journal of the American Society for Information Science **36**(1) (1985) 15–27
3. Viljanen, K., Tuominen, J., Hyvönen, E.: Publishing and using ontologies as mashup services. In: Proceedings of the 4th Workshop on Scripting for the Semantic Web (SFSW 2008), 5th European Semantic Web Conference 2008 (ESWC 2008), Tenerife, Spain (June 1-5 2008)
4. Viljanen, K., Tuominen, J., Hyvönen, E.: Ontology libraries for production use: The Finnish ontology library service ONKI. In: Proceedings of the 6th European Semantic Web Conference (ESWC 2009). (May 31 - June 4 2009)
5. Hyvönen, E., Viljanen, K., Tuominen, J., Seppälä, K.: Building a national semantic web ontology and ontology service infrastructure—the FinnONTO approach. In: Proceedings of the 5th European Semantic Web Conference (ESWC 2008). (June 1-5 2008)
6. Tudhope, D., Alani, H., Jones, C.: Augmenting thesaurus relationships: Possibilities for retrieval. Journal of Digital Information **1**(8) (2001)
7. Hollink, L., Schreiber, G., Wielinga, B.: Patterns of semantic relations to improve image content search. Journal of Web Semantics **5**(3) (2007) 195–203
8. Fu, G., Jones, C.B., Abdelmoty, A.I.: Ontology-based spatial query expansion in information retrieval. In: In Lecture Notes in Computer Science, Volume 3761, On the Move to Meaningful Internet Systems: ODBASE 2005. (2005) 1466–1482
9. Kauppinen, T., Hyvönen, E.: Modeling and reasoning about changes in ontology time series. In Kishore, R., Ramesh, R., Sharman, R., eds.: Ontologies: A Handbook of Principles, Concepts and Applications in Information Systems. Integrated Series in Information Systems, New York, NY, Springer-Verlag, New York (NY) (January 15 2007) 319–338
10. Jones, C., Abdelmoty, A., Fu, G.: Maintaining ontologies for geographical information retrieval on the web. Volume 2888., Sicily, Italy, Springer Verlag (November 2003) 934–951
11. Kauppinen, T., Väätäinen, J., Hyvönen, E.: Creating and using geospatial ontology time series in a semantic cultural heritage portal. In: S. Bechhofer et al.(Eds.): Proceedings of the 5th European Semantic Web Conference 2008 ESWC 2008, LNCS 5021, Tenerife, Spain. (June 1-5 2008) 110–123

[14] `http://www.lusto.fi`
[15] `http://smartmuseum.eu/`
[16] `http://www.seco.tkk.fi/projects/finnonto/`
[17] `http://www.seco.tkk.fi/projects/sw20/`

Publication IV

Matias Frosterus, Jouni Tuominen, Sini Pessala and Eero Hyvönen. Linked Open Ontology cloud: managing a system of interlinked cross-domain lightweight ontologies. ***International Journal of Metadata, Semantics and Ontologies*****, 10, 3, pages 189–201,**
DOI 10.1504/IJMSO.2015.073879, December 2015.

Noname manuscript No.
(will be inserted by the editor)

Linked Open Ontology Cloud – Managing a System of Interlinked Cross-domain Light-weight Ontologies

Matias Frosterus · Jouni Tuominen · Sini Pessala · Eero Hyvönen

the date of receipt and acceptance should be inserted later

Abstract Traditionally the structure of the controlled vocabularies used for annotation can be utilised for reasoning for information retrieval. However, this can be problematic when applied in the Linked Data context. Linked data typically comes from different organisations and domains with mutually incompatible vocabularies without explicit links between them resulting in data silos. This paper argues that to solve the problem one has to transform the annotation vocabularies into a *Linked Open Ontology* cloud. We present a method for transforming a set of legacy thesauri into a cloud of interlinked ontologies while ensuring the validity of the transitive subclass relations and the means for maintaining the system when component ontologies are updated. Our approach has been used and evaluated in practice building a cloud called KOKO of sixteen ontologies, with a total of 47,000 concepts. KOKO has been published as an ontology service and is in use in various organisations for both data indexing and semantic search.

Keywords Light-weight Ontologies · Linked Open Ontology Cloud · Ontology change propagation · SKOS · Ontologisation · Cross-domain interoperability

1 Introduction

Libraries, archives, museums, and other organisations have been using classifications and thesauri for content indexing (annotation) for a long time. There exist vast amounts of high-quality annotations describing vast amounts of documents and other objects but these metadata descriptions are often divided into silos without machine-traversable links between the values used in the metadata. Semantic interoperability between heterogeneous *cross-domain* data repositories of distributed content providers has become the critical challenge for the Semantic Web and Linked Data [13]. Enabling different thesauri and vocabularies to link to one another in meaningful ways would allow different organisations to benefit from each others' work by enriching a common pool of linked knowledge [15].

1.1 Why a Linked Open Ontology Cloud?

The Linked Data movement[1] has focused its efforts on building cross-domain interoperability by creating and using (typically) `owl:sameAs` mappings between the entities (e.g. places, persons) in the *datasets* of the Linked Open Data (LOD) cloud. However, when linking metadata, not only the data entities but also ontologies used in describing them need to be interlinked for interoperability. This calls for more refined ontology alignment techniques [8] maintaining the integrity of the concept hierarchies.

This paper focuses on aligning light-weight domain ontologies intended for metadata annotations. A light-weight ontology in our terminology is a hierarchy of concepts with subsumption, partitive, and associative relations like in a traditional thesaurus [2], and can be represented using RDFS[2], simple OWL[3] constructs, or SKOS[4].

The research hypothesis of this paper is that the LOD cloud could be complemented by developing one or more light-weight "Linked Open Ontology" (LOO) clouds. The idea of LOO is to provide a shared cross-domain ontology for data annotations based on a set of interlinked domain

Semantic Computing Research Group (SeCo), Aalto University
P.O. Box 15500, 00076 Aalto, Finland
http://www.seco.tkk.fi/, firstname.lastname@aalto.fi

[1] http://linkeddata.org/
[2] http://www.w3.org/TR/rdf-schema/
[3] http://www.w3.org/standards/techs/owl#stds
[4] http://www.w3.org/TR/skos-reference/

ontologies. This idea is also complementary with the idea of "Linked Open Vocabularies"[5] that focus on mapping metadata schemas (e.g. Dublin Core, FOAF, and Bibo) onto each other. In our implementation of the idea, we created the LOO cloud from light-weight ontologies based on existing thesauri. If the resulting LOO cloud is then mapped to, e.g., DBpedia, legacy data annotated using the original thesauri can be linked to the LOD cloud quite easily.

Developing a LOO cloud is in many ways different from linking entities in datasets or elements in metadata schemas. A major difference is that in LOO the linked structure is used for reasoning based on the hierarchical subclass relation, the backbone of ontologies [31]. This fundamental task requires special consideration at the ontology boundaries as otherwise cross-domain reasoning and ontology-based query and document expansion [3, 17] in applications may fail [16].

For example, assume that the concept "Mirror" is present in a given ontology A of daily utensils and has the subclass "Make-up-mirror":

`a:Make-up-mirror rdfs:subClassOf a:Mirror.`

In a related ontology B of furniture, the class Mirror is used as a subclass of the class Furniture:

`b:Mirror rdfs:subClassOf b:Furniture.`

Without context, the concept of mirror looks like the same in both ontologies, i.e.

`a:Mirror owl:sameAs b:Mirror.`

Reasoning and query expansion works fine in A and B separately, but when using A and B linked together, expanding a search query for "furniture" would return falsely handheld make-up mirrors in addition to pieces of furniture. A larger context than the concept alone has to be considered when linking ontologies.

There are also other difficulties specific to developing a LOO cloud. For example, the principle of dividing a shared concept X into subclasses in different ontologies may be different. For example, "clothes" can be divided into subclasses based on the gender or the age of their wearers. Then, from a human perspective, X may have a confusing mixture of subclasses in the linked ontology, hampering its use in user interfaces (e.g. as a search facet). Addressing issues like these is hard to automate, and therefore LOO development in practice requires more coordinated collaboration between the developers of linked ontologies than when linking datasets of instances. By collaboration, better quality links can be created and various critical issues of linked data quality[6] can be addressed. Coordinated collaboration also facilitates larger scale ontology development, which prevents the creation of interoperability problems, and minimises redundant ontology development work in overlapping areas of ontologies.

Our approach therefore emphasises the systematic development process of a coherent aggregated ontology from a set of component ontologies that are maintained by different domain communities. This *proactive* idea of developing a larger ontology based on different domain ontologies [16] is different from traditional ontology mapping, where one takes a set of existing, independently developed ontologies and tries to map them together *afterwards*. In contrast to simply mapping individual ontologies together we take the mappings as an integral part of the ontologies. The ontologies are considered both as individual entities but also as integral parts of the cloud forming a greater whole. This aspect is taken into account at all levels of the development and publication of the ontologies, thus leading to a different process and underlying philosophy compared with the usual approach of linking independently developed ontologies.

1.2 A National Effort in Building a LOO Cloud

In order to test and evaluate these hypotheses in practice, a LOO cloud called KOKO of cross-domain ontologies (e.g. health, cultural heritage, agriculture, seafaring, government, defence) has been realised in Finland on a national level during the FinnONTO research project (2003–2012). Various libraries, archives, museums, and governmental actors have annotated vast amounts of documents, each with their own thesauri. The aim of the KOKO cloud is to maintain backwards compatibility with the existing annotations while allowing for better interoperability between different datasets. Thus, the system is based on transforming a set of legacy thesauri in use into light-weight ontologies and interlinking them with each other. Currently, KOKO encompasses sixteen ontologised thesauri with more to be integrated into the system in the future. The current version of KOKO is a harmonised global ontology of some 47,000 concepts aligned into a single hierarchy.

In the centre of the KOKO cloud is the General Finnish Ontology YSO, based on the General Finnish Thesaurus YSA[7], which has been developed since 1987 by the National Library of Finland and is used across various organisations in Finland. YSO provides the upper hierarchy, originally inspired by the ideas of "foundational ontologies", such as DOLCE [9], and shared "upper ontologies", such as IEEE SUMO [27]. The idea is to provide the LOO cloud with the common upper concepts needed in many domains. YSO is then complemented by more refined ontologies for specific domains. These component ontologies are developed by the experts responsible for the original thesauri preserving the domain know-how while allowing for asynchronous updating based on the resources of each organisation. From an end user's point of view, KOKO ontology is seen and used

[5] http://lov.okfn.org/dataset/lov/

[6] Cf. e.g. http://pedantic-web.org/fops.html

[7] http://vesa.lib.helsinki.fi/ysa/

as a single ontology without boundaries; for the ontology developers, domain boundaries are needed in order to divide the responsibilities of distributed ontology work based on domain expertise needed in different parts of the KOKO cloud.

KOKO ontology was originally published in the ontology library system ONKI[8] [36,37] operating as a living lab research environment. Over the years ONKI has been integrated into, e.g. several museums, libraries, and web portals. In 2013 the National Library of Finland launched a joint project with the Ministry of Finance and the Ministry of Education and Culture to build a permanent, national ontology service Finto[9] based on the ONKI system [35]. The National Library also took on the responsibility of coordinating further development of the KOKO cloud. Finto ontology service provides a centralised publication channel for the ontologies with common interfaces for accessing them in various applications.

1.3 Challenges in Developing a Linked Ontology Cloud

Several issues need to be taken into account when moving from developing individual thesauri into developing and maintaining a cloud of interlinked ontologies. In this paper, the following challenges are discussed.

- **Creation of ontologies.** How is an existing legacy thesaurus transformed into an ontology? How is a new ontology mapped into the linked ontology system? How are the URIs of the concepts formed?
- **Development of ontologies.** How are the overlapping parts of ontologies recognised in order to minimise the duplicate work of the ontology developers? How are the changes in one ontology communicated to other related ontologies? How are the errors and other quality issues recognised in a system of several ontologies?
- **Publication of ontologies.** How should the linked ontology system be presented to the end user in order to make it comprehensible? What kind of services for using the ontologies are needed for different user groups?

1.4 Structure of the Paper

This paper is divided into four main parts. Section 2 presents a model for creating and updating a linked ontology cloud and lists a set of seven principles guiding the process. The following Sections 3–5 describe the development cycle of the cloud in more detail with an emphasis on how a single domain ontology is handled by the process. Throughout, the KOKO cloud provides an illustrative use case on how the process has been applied to practice. Since the traditional, comparative evaluation of the process is difficult, the extensive application to practice has been used as a proof of concept with the process being adjusted based on real-life experiences in accordance with the principles of action research [5]. The main user groups for this have been the ontology developers on the one hand, and the systems that KOKO has been integrated into on the other hand. In the final Section 6, related work is presented and the contributions of the paper are summarised.

2 A Model for Managing a Linked Ontology Cloud

Our ontology development work started by a field study on how thesauri in use are actually developed. The result was that thesauri are typically developed by independent expert groups focusing on concepts in their own domains of interest, with little collaboration between the groups. The situation seems to be more or less similar in other countries, too. Obviously, this model leads to redundant work in developing overlapping areas of thesauri and, at the same time, to interoperability problems between the thesauri, since different parties define their concepts without considering each others' choices. To address these problems we propose a more coordinated collaborative model for developing a linked ontology cloud, which is depicted in Fig. 1. Note that in the following discussion the bolded numbers in parentheses refer to the numbered parts in the figure.

2.1 Ontology Creation Phase

First, existing thesauri (**1**) and ontologies (**2**) are selected for building blocks of the ontology cloud. A thesaurus is converted into RDF format using a shared ontology schema (**3**) and aligned with a general upper ontology (GUO) (**4**). Aligning domain ontologies with a GUO forms the basis for interoperability by providing a complete concept hierarchy and is much easier to maintain than direct, pair-wise mappings between domain ontologies [12]. This idea was suggested, e.g. by the IEEE Standard Upper Ontology (SUO)[10] working group.

The alignment can be done in a semi-automatic fashion by first generating equivalency mappings automatically and then correcting them manually. In addition to equivalency mappings, subsumption and partitive relations might be used in the case of light-weight ontologies. For example, the concept "antique furniture" in a museum domain ontology may be aligned with the concept "furniture" in the upper ontology using a subclass-of relation. In order to create a complete, fully connected linked ontology, all concepts

[8] http://onki.fi/
[9] http://finto.fi/en/
[10] http://suo.ieee.org/

of a domain ontology should be mapped to the GUO either directly or through other concepts in the domain ontology. The transformation is discussed in more detail in [16]. In our case study, a natural basis for a GUO was the General Finnish Thesaurus YSA that was transformed into the General Finnish Ontology YSO. already as a reference in many specialised thesauri.

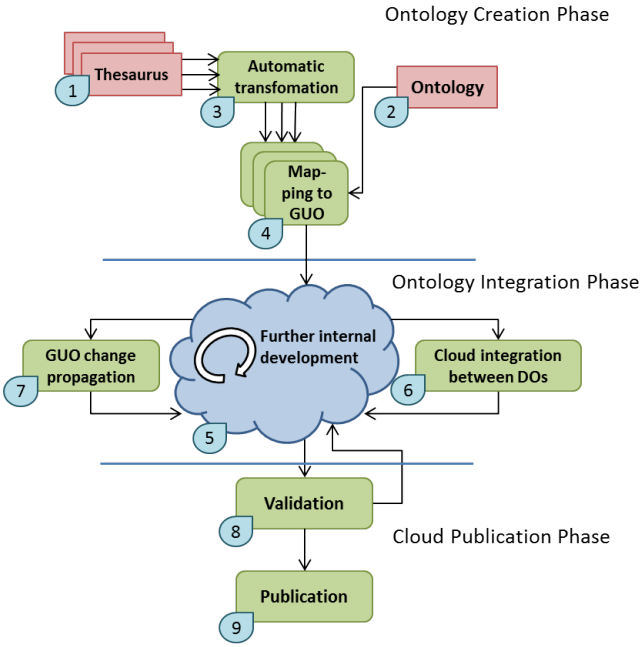


Fig. 1 The model of Linked Ontology Cloud formation and management

2.2 Ontology Integration Phase

The ontology integration phase begins after the domain ontologies have been aligned with the GUO (**5**). There may be mutually overlapping parts between the domain ontologies because the alignments were made between the domain ontologies and the GUO only. To facilitate the integration of domain ontologies (DO) (**6**), processes, and tools for discovering overlapping parts of the ontologies are needed. Based on the analysis, it is possible to eliminate redundant development work by deciding between domain ontology developers which ontologies maintain which overlapping parts.

Changes in the GUO create pressure for the domain ontologies to be updated accordingly to ensure the consistency of the cloud (**7**). For example, in our case study system, some 2,000 changes are made annually in the upper ontology YSO. The changes should be taken into account in the development process of the domain ontologies. Fortunately, not all changes in the GUO are relevant to all domain ontologies, but only those related to them via equivalency, subsumption or partitive relations.

2.3 Cloud Publication Phase

When a development cycle of an ontology cloud has been completed, its logical consistency and other quality aspects should be validated (**8**) making sure that the resulting ontology adheres to the constraints of the properties and classes used. Some of the problems encountered can be fixed automatically [34] (e.g. overlaps in disjoint semantic relations and cycles in the concept hierarchy) but they should nonetheless be communicated to the developers. However, automatic validation has difficulty in finding problems on the semantic level, which usually requires manual checking. Finally, the ontology cloud can be published as services for humans and machines, e.g. via user interfaces, APIs, and downloadable files (**9**).

2.4 Principles for Building a Linked Ontology Cloud

In our case study, the KOKO cloud based on the sixteen ontologised thesauri presented in Table 1 was created. Below, the lessons learned during the work are summarised as a seven-point list of practical building principles. We consider the proposed principles novel, as we are not aware of previous ontology design patterns focused on managing an ontology cloud as a whole.

I The ontology cloud consists of one general upper ontology and several domain ontologies that are linked to the upper ontology with subsumption, equivalency, associative and partitive relations.
Reason: This means that the domain ontologies do not have to be linked to each other pairwise, as the upper ontology acts as semantic glue for joining all the ontologies together. Shared concepts are included in the upper ontology. The idea is to minimise the links between the domain ontologies, which simplifies their development since a given developer needs only concern herself with her own domain ontology and the GUO.

II Every concept in a domain ontology has a subsumption or equivalency relation to a concept in the GUO or a subsumption relation to a concept in the same domain ontology.
Reason: This means that every concept in a domain ontology needs to be able to trace a subsumption relation to a concept in the general upper ontology. This ensures a consistent concept hierarchy for the whole cloud and that domain ontologies can not define new top-level concepts.

Name of the ontology	Domain	Number of concepts
YSO	General upper ontology (GUO)	27,200
AFO	Agriculture and forestry	7,000
JUHO	Government	6,300
KAUNO	Literature	5,000
KITO	Literary research	850
KTO	Linguistics	900
KULO	Cultural research	1,500
LIITO	Economics	3,000
MAO	Museum artefacts	6,800
MERO	Seafaring	1,300
MUSO	Music	1,000
PUHO	Military	2,000
TAO	Design	3,000
TERO	Health	6,500
TSR	Working and employment	5,100
VALO	Photography	2,000

Table 1 The ontologies comprising the LOO cloud KOKO

III If a concept in a domain ontology has an equivalency to a concept in the GUO, it may not have broader concepts in the domain ontology, which lack an equivalency relation to a concept in the GUO.
Reason: This is needed to avoid having dependencies from the GUO towards a domain ontology by forbidding domain ontologies from introducing broader concepts to concepts in the GUO (through inference over equivalency relation). Otherwise domain ontology developers could propagate contradictions or unwanted concept (re)definitions into the GUO, especially in cases where more than one domain ontology is involved.

IV The domain ontologies are focused on a clearly bounded domain and are as self-contained as possible.
Reason: This allows the domain ontology developer to concentrate on the area of her expertise. This also minimises dependencies between domain ontologies and facilitates ontology development work.

V A concept in a domain ontology may not have an equivalency to a concept in another domain ontology. A concept in a domain ontology may have an associative relation or have a broader concept in another domain ontology at the discretion of the developers.
Reason: Dependencies between domain ontologies cannot always be avoided due to inherent relations across the borders of different domains. This means that the developers need to monitor the changes to the other domain ontology but this is allowed since it does not affect the other domain ontology directly. The use of broader and associative relations extends the target domain ontology but does not affect its semantics (strongly), whereas the equivalency relation would possibly introduce new broader concepts to concepts in the target ontology and thus redefine their semantics.

VI The GUO and the domain ontologies use a shared ontology schema and are based on domain-independent standards.
Reason: Thus the ontologies can be easily merged and used together with various data sources outside of those directly linked to the cloud, using standard software, e.g. SKOS or RDFS/OWL tools.

VII The resulting ontology cloud should be logically consistent, e.g. by ensuring the integrity of the concept subsumption hierarchy over ontology boundaries (since transitivity is assumed).
Reason: This allows for reasoning and query expansion over the whole cloud.

3 Creating a Thesaurus-based Ontology for the Linked Ontology Cloud

The creation of a cloud of linked ontologies begins with the formation of the general upper ontology. This ontology provides a completely connected hierarchy of general concepts including the topmost division, e.g. between abstract, endurant and perdurant concepts [9]. This forms the basic structure for the domain ontologies to map into, thus saving them from having to repeat the higher parts of the hierarchy.

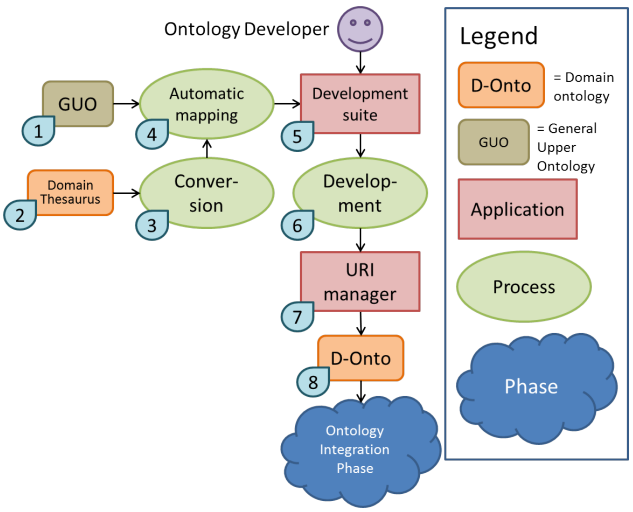


Fig. 2 Ontology Creation Phase

Fig. 2 depicts the process of creating a new domain ontology for the cloud of linked ontologies, which begins with the conversion (**3**) of the domain thesaurus (**2**) from a legacy

format into RDF, typically SKOS. This is often a straightforward operation but in some cases the original thesaurus can include relations that are not easy to convert to SKOS. In these cases it can be a good idea to retain the original relations in the form of a temporary predicate which can then be harmonised in the publication phase of the process.

In the FinnONTO case, we used ad-hoc scripts for the transformation since the domain ontologies were in different formats. We also transformed them originally into a custom OWL-based format since at the beginning of the project SKOS had not yet been established as a standard way of representing light-weight RDF ontologies. We continued this practice in part because of existing tools, such as Protégé[11], but also because some thesauri used relations not present in SKOS. An example of this was the Agriforest[12] thesaurus of agriculture and forestry, which uses different relations to differentiate between names of concepts that were derived from different sources. These the were preserved for the benefit of the ontology developers but were then combined into a common predicate for publication.

The next step of the process is the automatic mapping (**4**) to the GUO (**1**). This can be done roughly through string comparison between labels or, alternatively, by also utilising the structure of the ontologies, depending on the nature of the original thesaurus. Since different thesauri may have different labelling conventions (for example plural vs. singular forms) some sort of stemming or lemmatisation may also be needed for the label comparison. The result is a preliminary mapping which then needs to be checked manually in the next part (**5** and **6**) by a domain expert ontologist since the aim is to produce as good and reliable a result as possible. Aside from checking the mapping, the human development part also entails the ontologisation work proper, which needs to be done when changing vaguely defined terms into more precise ontology concepts [16,21]. In many cases, a term in the original thesaurus becomes several concepts in the ontology depending on whether the term has multiple meanings. When a single term corresponds to several concepts, we have added a qualifier in parentheses to the preferred label for each concept in order to make it easier to differentiate between them. For example, the ambiguous keyword ”child” could be split into concepts ”child (age)” and ”child (family relation)”.

In the FinnONTO project, we did the preliminary mapping between a domain ontology and the GUO by label comparison using lemmatisation and custom scripts, while most of the actual ontologisation work was done by the experts from the organisations that maintained the thesauri. Now there exist specialised ontology matching tools, such as AgreementMakerLight[13], which could be utilised for the initial mapping.

Finally, the URI scheme of the ontology needs to be considered. A good starting point is the URI design principles presented by W3C in *Cool URIs for the Semantic* Web[14]. Human-readable meaning-bearing URIs pose difficulties in that they are language-dependent and, most importantly, in the case where further development ends up changing the label of the concept, the persistent URI can not keep up with the changes, thus gradually leading to the degeneration of the mapping between the labels and URIs. Since we are building on thesauri that have been in active use and development for long, we must also consider the long term effects of our current decisions. Therefore, we decided to use language- and meaning-neutral URIs where the local name consisting of a letter and a string of numbers (e.g. `p3612`).

Since ontologies evolve with new concepts and URIs introduced and old ones possibly deleted, a system (**7**) is needed for tracking the use of the URIs. In FinnONTO, we created our own tool for this called Purify[15]. Purify harmonises all local names in a given namespace to the letter followed by a number format. The tool keeps a log of the mappings between the possible temporary URIs used during the early stages of the development since human-readable URIs were found to be useful at the beginning of the ontologisation. Finally, Purify makes sure that no two resources get the same URI and that if the URI generation needs to be repeated, the result will be the same every time.

With the first version of a domain ontology completed (**8**) and successfully mapped to the GUO, the next step is to integrate it into the ontology cloud proper. Table 1 lists all the ontologies completed for our LOO cloud KOKO at the moment. The name of the ontology is followed by a description of its domain and the number of concepts.

4 Maintaining a Cloud of Interlinked Ontologies

There are two main concerns when integrating and managing domain ontologies in a cloud: how to manage the domain ontologies with respect to one another and how to manage their relation to GUO. This phase is depicted in Fig. 3, continuing from Fig. 2, and is explained in depth below.

4.1 Avoiding Overlap Between Domain Ontologies

The Principles IV and V presented in Section 2 posit that in order to keep relations between domain ontologies to a minimum, the domains covered need to be precisely set. With

[11] http://protege.stanford.edu/
[12] http://www-db.helsinki.fi/agri/agrisanasto/Welcome_eng.html
[13] https://github.com/AgreementMakerLight
[14] http://www.w3.org/TR/cooluris/
[15] http://puri.onki.fi/info/

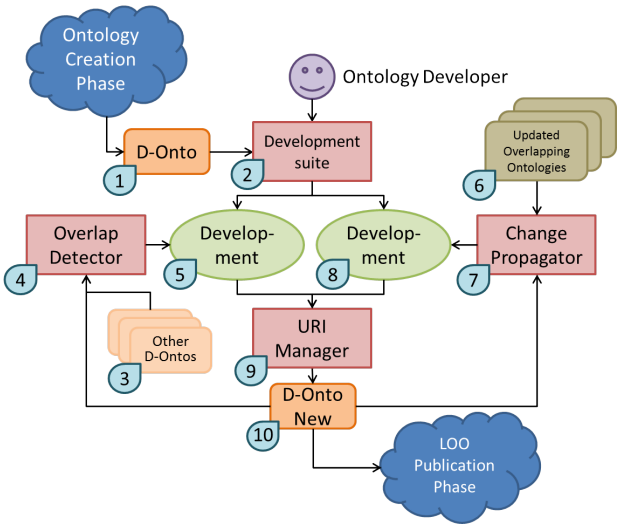


Fig. 3 Ontology Integration Phase

minimal links between domain ontologies, they can be developed independently from one another, but it also means that the curation of the ontological domains is an on-going process. Overlaps can come into being especially in the border areas of two domains where a given set of concepts can not be unequivocally said to belong to either one of two different ontology domains.

In Fig. 3, we can see that the process of discovering these overlaps between a given domain ontology (**1**) and the rest of the domain ontologies in the cloud (**3**) makes use of a tool (**4**) to help the domain ontology developers in finding and even analysing the overlaps. This tool can be based on string matching similar labels and reporting on the potentially overlapping concepts, but it can also make use of the ontological structure and relations to find overlapping concepts [8].

Once an overlap between two or more domain ontologies has been discovered, a dialogue (**5**) should be started between the developers of those ontologies. Its possible end results are as follows:

a) The concepts that overlap are really the same concept and the developers of the domain ontologies and the GUO can agree that the concept is general enough to be included in the GUO. In this case, the concept can be removed from the domain ontologies.
b) The concepts that overlap are really the same concept but the developers can agree that the concept is not needed in both (or all) of the domain ontologies and agree on a single domain ontology that should host the concept in question and handle changes and development further on that concept and its subconcepts. This is most common in situations where the concept is clearly in the domain of one ontology and only included in the other due to historical reasons, for example.
c) The concepts that overlap are really the same concept and the ontology developers wish for it to remain present in both (or all) of the ontologies. A note should be made that if the concept is changed or developed further, the other developers should be informed as needed.
d) The concepts that overlap are actually different concepts but might share a preferred label. In this case, the labels should be differentiated from one another if possible so as to avoid confusion when the ontologies are used together.

Great care should be exercised when choosing option c) since it clashes with Principle V and can easily lead to increasing complexity in development. This should be avoided as much as possible, since one of the main goals of the presented system is to make the asynchronous development of ontologies possible and having the same concept in two domain ontologies means that both sets of developers need to agree on possible further development.

In our case study, we implemented a tool called KOAN, based on simple label matching, for finding ontology overlaps since the light-weight ontologies do not offer much structure to use as a basis for the discovery task. Furthermore, identical labels pose the most difficulty for the annotators using the ontology since they would need to decide between different concepts with the same names based on their place in the hierarchy. Having concepts with different labels that end up being the same is much less of a problem since, when the mistake is discovered, it is relatively easy to combine the annotations made using the duplicated concepts.

When applied to KOKO, KOAN found lots of overlapping concepts even between ontologies of seemingly very distinct domains. Table 2 shows some of the comparison results between ontologies by showing the per cent amount of overlap. For instance, from the first row, second cell, we can see that the agriculture and forestry domain ontology AFO contains 8 % of the concepts of the government domain ontology JUHO. Comparing with Table 1 we can see that even ontologies from seemingly very distinct domains can have a lot of overlap between them. Maintaining the overlaps in multiple places at the same time creates a lot of unnecessary work and can lead to inconsistencies in the transitive relations. Our aim is to implement a systematic process for the elimination of these overlaps.

In addition to overlapping concepts, a domain ontology may have a concept with an associative relation or a broader concept in another domain ontology because cross-domain relations cannot always be avoided. For example, in our use case, the museum domain ontology MAO contains the concept ”catapults”, which is a subclass of the concept “weapons” in GUO. On other hand, the military domain ontology PUHO has the concept “single-shot weapons”, which could be used as the superclass of “catapults” for more refined semantics. However, using relations between domain ontologies may

Ontology	AFO	JUHO	KAUNO	MERO	TERO
AFO		8%	2%	3%	25%
JUHO	7%		16%	5%	40%
KAUNO	2%	12%		1%	28%
MERO	0%	1%	0%		2%
TERO	23%	41%	36%	13%	

Table 2 Number of overlapping concepts in five domain ontologies of KOKO

be a justification of moving such cross-domain concepts ("single-shot weapons" in the example) into GUO in order to minimise dependencies between domain ontologies.

4.2 Keeping the Domain Ontologies up to Date Regarding GUO

The second part of the cloud management process, as depicted on the right hand side of Fig. 3, is the handling of the changes in the upper ontology. As an implication of Principle II, the domain ontologies react to the changes in the upper ontology. In other words, when the GUO developers release a new version (**6**), the domain ontologies need to be potentially updated.

The structure of the cloud aims to allow for the asynchronous updating of the domain ontologies, which can result in long intervals between the updates of a given domain ontology. Additionally, the ontologies are often developed in different organisations, and using separate ontology editors creates a challenge in how to communicate the changes between the developers. The two main approaches to conveying the changes are push and pull synchronisation [6]. The push version propagates the changes in one ontology immediately to the other ontologies, whereas in the pull version the change listing is requested when needed by the ontology developer. The push approach is good for situations where the ontologies are updated frequently, so that the ontology developer can quickly ensure the consistency of the ontology after the changes. However, this approach is challenging if the ontology reacting to changes is update infrequently due to, e.g., lack of resources. Then it would be preferable that the changes could be acquired when needed and not propagated immediately. A long update interval also means that the amount of changes can build up over time and when the update process of the domain ontology is started, the number of changes that need to be checked can be in the thousands.

In order to ease the work of the domain ontology developer, a tool (**7**) needs to be used for propagating the changes. It would also be beneficial to order or categorise the changes of the GUO somehow according to their relevance to the domain ontology in question. A set of criteria for estimating relevance should take into account the differences in ontologies and the relations between them as well as the likely changes that are going to occur in development.

In the FinnONTO project, we created a change propagation tool MUTU and a set of accompanying relevance criteria as described in [29]. The basic idea is that a change in the GUO is likely to be relevant to domain ontology developers if it concerns a concept that has been directly linked to from the domain ontology (a connecting concept). If a concept in the GUO has been marked as equivalent or as a superclass to a concept in the domain ontology, any change to it is likely to be of interest to the domain ontology developers. Furthermore, if the concept hierarchy of an ancestor changes for a connecting concept in GUO, this is likely to be relevant to the domain ontology developer due to the transitive nature of the relation. If a concept is removed, rare though that is, that is deemed as interesting due to the fact that this concept could have been in use by the domain ontology users but has not been duplicated to the ontology itself. Finally, if a concept has been added that has the same label as a concept already existing in the domain ontology, this is always relevant.

MUTU simply finds out the changes between the new version of the GUO and the one that was used for the mapping of the domain ontology, lists these changes according to the type of the change and relevance, and adds helper classes to the development version of the domain ontology for grouping the concepts in the ontology editor suite that need to be checked by the developer (**8**). MUTU also allows the domain ontology developer to configure it according to her needs based on, e.g., a specific priority of languages or on blocking changes from certain properties deemed irrelevant to the domain ontology development.

5 Publishing a Linked Open Ontology Cloud

Once the individual component ontologies have been developed and the links between them have been curated, the ontology cloud needs to be published. The aim is to provide the users with a single, unified whole that can be used essentially as a single ontology like any other. The process of publishing a Linked Open Ontology Cloud is depicted in Fig. 4.

5.1 Validation, Merging and Publication

The publication phase starts when a new version of a domain ontology (**1**) is ready and the developer wants to publish it in the LOO cloud. The first step is to clean up the ontology for publication (**2**), by removing structures needed

only in the development of the ontology. For example, temporary concepts that are used for grouping concepts that are under development and editorial notes for internal use by the ontology developers can be removed. Similarly, temporary ontology-specific predicates should be converted to the common schema.

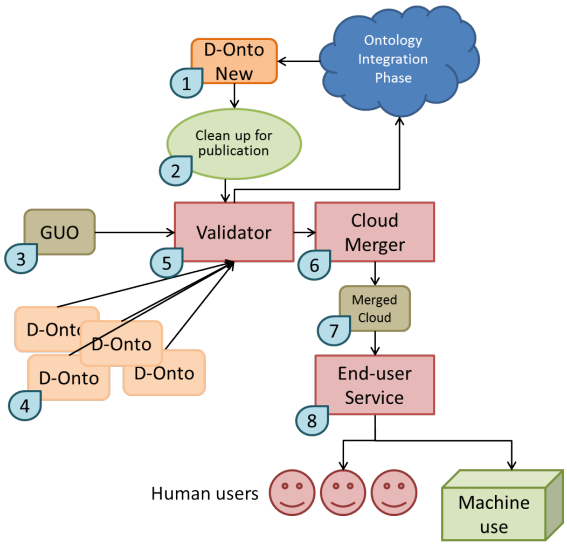


Fig. 4 Cloud Publication Phase

In the FinnONTO case, the domain ontologies were developed in a Protégé project with the general upper ontology included. To help the ontology developer to focus on the concepts of the domain ontology, the topmost concepts of the domain ontology (the ones that do not have subclass-of relation to a concept in the domain ontology) are connected with subclass-of relation to a temporary concept acting as a root concept for the ontology. This root concept is removed in the clean up process as the goal is to present the resulting ontology cloud to the end users as a single complete hierarchy with a single root concept.

Before the new domain ontology version is merged into the cloud it is validated (**5**) for compliance with Principle VII, e.g. by checking the logical consistency and spotting the violations of best practices. The results of the validation are communicated to the ontology developer who can then fix the problems. A validator may also fix some of the problems automatically, e.g. by removing cycles in the concept hierarchy.

For the validation of the ontologies we are using the Skosify tool [34], which converts RDFS/OWL/SKOS vocabularies into proper SKOS format. Overlaps in disjoint semantic relations (`skos:related` and `skos:broaderTransitive`) are checked and inconsistencies are corrected automatically by removing `skos:related` relations in problematic cases. In cases where a concept has more than one `skos:prefLabel` per language one of the labels is arbitrarily selected for use as `skos:prefLabel` while the others are simply converted into `skos:altLabels`. A check is performed in order to detect overlaps in disjoint label properties (`skos:prefLabel`, `skos:altLabel`, and `skos:hiddenLabel`) and the superfluous ones are removed. As a best practice, the cycles in the concept hierarchy are removed, and finally extra whitespaces are removed from concept labels. The violations are also reported to the ontology developer so he may fix the problems in a controlled way instead of relying on the automatic fixing procedure. We also used Skosify to convert the ontologies from the custom OWL format to SKOS for publication. SKOS was chosen due to the amount of tools available and for its suitability for our use case of light-weight ontologies.

After the validation, the new version of the domain ontology is merged (**6**) with the GUO (**3**) and the other domain ontologies (**4**). The idea is to build a single representation of the linked ontology cloud (**7**) by merging equivalent concepts and giving them a single URI. The differing ontology schemas are harmonised so that the end user does not get confused with the ontology-specific structures and naming conventions.

When annotating using the linked ontology cloud, the annotator should be making choices between concepts as opposed to ontologies. To this end, the cloud is merged (**6**) into a single whole (**7**) by taking all the concepts linked with `skos:exactMatch` relations between them (marked as equivalent) and combining them. In case a clearer division between the development version of individual ontologies and the published version of the cohesive cloud needs to be made, new URIs can also be assigned to the concepts of the cloud. Here care needs to be taken in order to ensure that the same concepts always map to the same URIs even in cases where that concept might at first originate from the upper ontology and later have been moved to a domain ontology or vice versa.

In FinnONTO, we gave all the concepts in the KOKO cloud new URIs in a new namespace. The combination of the ontologies listed in Table 1 resulted in a combined cloud of some 47,000 concepts. In order to allow the end user to select only from a single domain, we assigned domain ontology specific concept types as subclasses of `skos:Concept`.

Once the ontology cloud is merged, it is ready for publication (**8**). In accordance with the "open" part of the name LOO, the cloud is published as Linked Data [13], so that individual concept URIs can be referenced from various datasets and URIs are resolvable to information about the concepts in human- and machine-readable forms. The LOO cloud is published for humans via user interfaces for searching, browsing and visualising the ontologies, and as widgets for integrating ontologies into applications. For machine use, additional REST and Web Service APIs are provided to facilitate

even deeper integration of the ontologies. Finally, the ontology cloud is published as a SPARQL endpoint enabling ad hoc queries for more complex needs.

These services were provided by the ONKI Ontology Service [37], part of which was further developed by the National Library of Finland in the form of Finto thesaurus and ontology service. These services act as repositories for vocabularies, thesauri, and ontologies, providing support for publishing, finding and accessing them. The main user groups of ONKI and Finto are ontology developers, content indexers and information searchers. Ontology developers need a way to visualise the ontology they are developing, and especially in the context of the LOO cloud, where the domain ontology developers map their ontologies only to the general upper ontology, there is a need for getting an overview of the whole cloud. This way the developers can see how their domain ontologies are situated in the LOO cloud and discover possible overlaps with other domain ontologies.

Content indexers and information searchers need ways of finding ontologies and concepts suited for their needs. The Linked Open Ontology Cloud approach eliminates the need for finding ontologies and making selections between them, as one ontology system covering all domains of life aims to fulfill the needs of everyone. For finding suitable concepts, ONKI/Finto service provides an ontology browser which visualises the ontologies as a tree hierarchy and shows other relations between concepts. The user may use autocompletion search for finding concepts based on their labels. In addition to dereferenceable URIs, machines are served with REST API, and a SPARQL endpoint[16]. The general ontology service software powering Finto is developed currently by the National Library of Finland as an open source project Skosmos[17] and can be freely used to set up a similar service. The ontology cloud is published under a permissive Creative Commons Attribution license[18].

5.2 Use Cases

During the FinnONTO research project, the adoption of KOKO was hampered by the uncertainty regarding its future. With the ontology service project of the National Library of Finland, the development and publication of KOKO was given sustainable governmental resources thus securing its future. However, due to the length of the process of securing funding, the wide-spread adoption of KOKO is only beginning.

KOKO and ONKI/Finto have been in daily use in many museums, such as the Espoo City Museum, where it has been integrated into the collection management system Kauko. The first large scale system using KOKO was the semantic cultural heritage portal CultureSampo[19] [25] aggregating contents from various data sources. KOKO is a basis of the deployed BookSampo[20] [24] literature portal of the Finnish public libraries with some 60,000 monthly end users on the Web.

At the moment, Finnish archives are integrating KOKO ontologies via Finto into their common search registry service for metadata on both digital as well as conventional documents with the AHAA project [19]. A recent company pilot user of KOKO has been the Swedish language division of the national public service broadcasting company of Finland, Svenska YLE. They have used KOKO in the annotation of their web content and have complemented it with Freebase[21] for instance references, such as people and organisations. The pilot has been a success and they are considering adopting the system in the Finnish part of YLE as well as their archive.

The multi-disciplinary nature of KOKO means that it is especially suited to organisations that deal with cross-domain contents such as media organisations, museums, libraries, and archives. Furthermore, using the concepts from domain ontologies links the content to more in-depth data from the specialist organisations that originally developed the domain ontologies for their data annotation needs.

6 Conclusions and Discussion

We described a process for creating and managing a Linked Open Ontology Cloud, based on existing legacy thesauri. To conclude, we compare our approach to related work, summarise our contributions, and suggest future work.

6.1 Related Work

Our work on linking ontological concepts used in annotations complements the idea of the Linked Open Data cloud, where emphasis is on data entity linking using (typically) `owl:sameAs` properties [11]. We also complement the work on Linked Open Vocabularies (LOV)[22], which focuses on metadata schema linking based on property hierarchies rather than linking domain ontologies. Linking the data through ontologies allows additional interoperability due to the inferenced knowledge gained through the shared ontology semantics [14]. Our approach follows this principle and provides a framework for interlinked ontology cloud develop-

[16] http://api.finto.fi/sparql; The KOKO cloud is available in the named graph http://www.yso.fi/onto/koko/.
[17] http://github.com/NatLibFi/Skosmos/
[18] http://creativecommons.org/licenses/by/3.0/
[19] http://www.kulttuurisampo.fi/
[20] http://www.kirjasampo.fi/
[21] https://www.freebase.com/
[22] http://lov.okfn.org/dataset/lov/

ment. Backwards compatibility with existing annotations is retained by basing the ontologies on legacy thesauri in use.

Another example of highly linked ontologies can be found in BioPortal[23], an ontology repository that has features supporting collaborative (inter-)ontology development. In addition to uploading new ontologies to BioPortal, users can also create and upload mappings between the concepts of different ontologies [28]. The mappings can be used to bridge overlapping ontologies or to extend general ontologies with specialised ones. Users can comment and create discussion threads on ontologies, their parts, and mappings, thus supporting a collaborative and open inter-ontology development process. However, a set of mapped ontologies does not appear as a single whole to the user though one can browse the ontologies by following the mappings between concepts. Moreover, the mappings are not utilised in the ontology development process to propagate the changes of an (upper) ontology to ontologies extending it, as they are in our LOO model.

In order to move from a group of ontologies into a coherent ontology system, the ontologies need to be reconciled. Ontology reconciliation [12] is a broad term, covering ontology merging, alignment, and integration. Most of the reconciliation methods are automatic or semi-automatic, which can lead to lower quality, especially if the ontologies were originally expert-made [7]. Our approach emphasises the importance of manual development work in the reconciliation process.

In ontology modularisation [1], ontologies are divided into smaller interlinked parts to facilitate distributed development and re-use. Our approach merges ontologies based on existing thesauri to form a Linked Open Ontology Cloud, while the development of the individual ontologies continues in a modularised way. Furthermore, we present the resulting interlinked cloud as single whole so that the end users do not have to make selections between ontologies.

There have been several previous efforts on building a general upper ontology [26] that can be used as a foundational basis for domain ontologies. Some of the upper ontologies have been developed from scratch, while, e.g. the Suggested Upper Merged Ontology SUMO [27] was created by merging existing ontologies. We used the General Finnish Ontology YSO, transformed from an existing general thesaurus YSA, as the upper ontology in our Linked Open Ontology Cloud. In contrast to the previous work on upper ontologies, which focuses on the creation of an upper ontology, our work emphasises the model for managing the domain ontologies as part of the ontology cloud and keeping them up to date and synchronised with the changes of the upper ontology.

Related to the principles of forming and managing the ontology cloud presented in this paper, similar guidelines have been presented in the context of the OBO Foundry initiative [30]. However, in OBO Foundry the focus is on coordinating the development of different ontologies under shared principles, but not on merging them into a single ontology cloud and the challenges therein. General, domain-independent ontology design principles have been proposed by several researchers [10, 38]. Our model uses their ideas as a foundation, e.g. by organising orthogonal concept domains into separate ontologies and supporting ontologies of different granularity levels in the form of a GUO and the domain ontologies extending it. The principles presented in this paper extend the general ontology design principles by covering modelling issues related to the management of changes in a linked ontology cloud.

Keeping domain ontologies up to date with the changes of the GUO is closely related to the topic of ontology evolution, which concentrates on addressing the changes in ontologies over time. Different change types have been listed in [33], whereas [20] considers more abstract change patterns constructed from atomic changes. According to [4] users would have liked to see explicitly when a concept created by them had been modified, indicating that not all the changes occurring in an ontology are equally relevant to all developers providing the impetus for our work on building a set of priorities for different changes. The detection of changes in an updated ontology can be done using logs [32] or by comparing two versions of the ontology [22], which was the approach we chose. Additionally, the extra challenges of distributed ontology development have been addressed in [20, 23, 18]. In our approach there is no assumption that all the ontology developers use the same ontology editor, as the support tools are implemented separately from the ontology development suite.

6.2 Contributions and Future Work

We presented the idea of the Linked Open Ontologies (LOO) for fostering data integration. LOO complements dataset entity linking in LOD and metadata schema linking in LOV. The realisation of the LOO cloud was achieved by facilitating an environment of interconnected but easily manageable set of cross-domain ontologies allowing for distributed ontology development. This can be more efficient since the mappings between ontologies can be re-used for different datasets.

Furthermore, we presented a detailed description of managing the overlaps between domain ontologies and the inconsistencies resulting from the asynchronous updating of the GUO compared with the domain ontologies. Finally, we implemented a LOO in practice with sixteen ontologies and a full set of tools for the development cycle from a set of thesauri to a fully realised Linked Open Ontology Cloud. Based on this work, we accompanied the LOO model with

[23] http://bioportal.bioontology.org/

a set of seven principles that guide the process of building and managing an ontology cloud. The principles are general enough to be applied to other ontology clouds in addition to the one we have built.

A central challenge in the linking process is in maintaining the integrity of the transitive relations across different ontologies and we have achieved this through the use of a general upper ontology acting as a central hub for the linking of various domain ontologies. We consider proactive linking as an integral part of the development process itself as opposed to simply mapping independently developed ontologies.

The model was piloted by implementing it into practice in extensive scale in various organisations encompassing many different domains. The model evolved based on lessons learned during the multi-year implementation and development process. Feedback was gathered from ontology developers as well from the systems integrating the linked ontology cloud. Based on the success of the prototype, the model is now being applied on a national scale in archives, libraries, and museums, as well as in ministries and other governmental agencies.

For our future research, we are focusing on building better tools and refining the processes for tracking and communicating the changes in the GUO to the domain ontology developers. To this end, we are also devising a more formal administrative process for the development and overall coordination of the KOKO cloud. Since it is a joint operation by over a dozen different organizations the particulars of the administration need to be both flexible yet well-defined.

References

1. S. B. Abbès, A. Scheuermann, T. Meilender, and M. d'Aquin. Characterizing modular ontologies. In *Proceedings of the 6th International Workshop on Modular Ontologies*. CEUR Workshop Proceedings, `http://CEUR-WS.org`, July 2012.
2. J. Aitchison, A. Gilchrist, and D. Bawden. *Thesaurus Construction and Use: A Practical Manual*. Aslib IMI, 2000.
3. J. Bhogal, A. Macfarlane, and P. Smith. A review of ontology based query expansion. *Information Processing & Management*, 43(4):866–886, 2007.
4. S. Braun, A. Schmidt, A. Walter, and V. Zacharias. The Ontology Maturing Approach to Collaborative and Work Integrated Ontology Development: Evaluation Results and Future Directions. In *Proceedings of the International Workshop on Emergent Semantics and Ontology Evolution at (ISWC 2007)*, pages 5–18, 2007.
5. F. Burstein and S. Gregor. The systems development or engineering approach to research in information systems: An action research perspective. In *Proceedings of the 10th Australasian Conference on Information Systems*, pages 122–134. Victoria University of Wellington, New Zealand, 1999.
6. P. Deolasee, A. Katkar, A. Panchbudhe, K. Ramamritham, and P. Shenoy. Adaptive Push-Pull: Disseminating Dynamic Web Data. In *Proceedings of the 10th international conference on World Wide Web (WWW 2001)*, pages 265–274, 2001.
7. Z. El Jerroudi and J. Ziegler. iMERGE: Interactive Ontology Merging. In *Proceedings of the International Conference on Knowledge Engineering and Knowledge Management (EKAW 2008)*, pages 52–56, 2008.
8. J. Euzenat and P. Shvaiko. *Ontology Matching*. Springer–Verlag, 2007.
9. A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider. Sweetening ontologies with dolce. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, EKAW '02, pages 166–181. Springer-Verlag, 2002.
10. T. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43(5-6):907–928, 1995.
11. H. Halpin, P. J. Hayes, J. P. McCusker, D. L. McGuinness, and H. S. Thompson. When owl:sameAs isn't the same: An analysis of identity in linked data. In P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, and B. Glimm, editors, *The Semantic Web – ISWC 2010*, volume 6496 of *Lecture Notes in Computer Science*, pages 305–320. Springer Berlin Heidelberg, 2010.
12. A. Hameed, A. Preece, and D. Sleeman. Ontology reconciliation. In S. Staab and R. Studer, editors, *Handbook on ontologies*, pages 231–250. Springer-Verlag, Germany, 2004.
13. T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space (1st edition)*. Morgan & Claypool, Palo Alto, California, 2011.
14. P. Hitzler and F. van Harmelen. A reasonable semantic web. *Semantic Web Journal*, 1(1-2):39–44, 2010.
15. E. Hyvönen. *Publishing and using cultural heritage linked data on the semantic web*. Morgan & Claypool, Palo Alto, California, October 2012.
16. E. Hyvönen, K. Viljanen, J. Tuominen, and K. Seppälä. Building a National Semantic Web Ontology and Ontology Service Infrastructure—The FinnONTO Approach. In *Proceedings of the European Semantic Web Conference (ESWC 2008)*, pages 95–109, 2008.
17. K. Järvelin, J. Kekäläinen, and T. Niemi. ExpansionTool: Concept-based query expansion and construction. *Information Retrieval*, 4(3/4):231–255, 2001.
18. E. Jiménez Ruiz, B. Cuenca Grau, I. Horrocks, and R. Berlanga. Supporting concurrent ontology development: Framework, algorithms and tool. *Data & Knowledge Engineering*, 70(1):146–164, 2011.
19. J. Kilkki, O. Hupaniittu, and P. Henttonen. Towards the new era of archival description – the Finnish approach. In *Proceedings of the International Council on Archives Congress*, August 2012.
20. M. Klein. *Change Management for Distributed Ontologies*. PhD thesis, Vrije Universiteit Amsterdam, 2004.
21. D. Kless, S. Milton, and E. Kazmierczak. Relationships and relata in ontologies and thesauri: Differences and similarities. *Applied Ontology*, 7(4):401–428, 2012.
22. K. Kozaki, E. Sunagawa, Y. Kitamura, and R. Mizoguchi. A framework for cooperative ontology construction based on dependency management of modules. In *Proceedings of the International Workshop on Emergent Semantics and Ontology Evolution (ISWC2007)*, pages 33–44, 2007.
23. A. Maedche, B. Motik, and L. Stojanovic. Managing multiple and distributed ontologies on the Semantic Web. *The International Journal on Very Large Data Bases (The VLDB Journal)*, 12(4):286–302, 2003.
24. E. Mäkelä, K. Hypén, and E. Hyvönen. BookSampo—lessons learned in creating a semantic portal for fiction literature. In *Proceedings of ISWC-2011, Bonn, Germany*. Springer–Verlag, 2011.
25. E. Mäkelä, T. Ruotsalo, and E. Hyvönen. How to deal with massively heterogeneous cultural heritage data—lessons learned in CultureSampo. 3(1), 2012.
26. V. Mascardi, V. Cordì, and P. Rosso. A comparison of upper ontologies. In *Proceedings of the 8th Workshop Dagli Oggetti*

agli Agenti: "Agenti e Industria: Applicazioni tecnologiche degli agenti software" (WOA 2007), pages 55–64, September 2007.
27. I. Niles and A. Pease. Towards a standard upper ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS 2001)*, pages 2–9. ACM, October 2001.
28. Natalya F. Noy, Nicholas Griffith, and Mark A. Musen. Collecting community-based mappings in an ontology repository. In *Proceedings of the 7th International Semantic Web Conference (ISWC 2008)*, pages 371–386. Springer–Verlag, October 2008.
29. S. Pessala, K. Seppälä, O. Suominen, M. Frosterus, J. Tuominen, and E. Hyvönen. MUTU: An analysis tool for maintaining a system of hierarchically linked ontologies. In *Proceedings of the Workshop on Ontologies come of Age Workshop (ISWC 2011)*, 2011.
30. B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, The OBI Consortium, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255, 2007.
31. S. Staab and R. Studer, editors. *Handbook on Ontologies (2nd Edition)*. Springer–Verlag, 2009.
32. L. Stojanovic, A. Maedche, B. Motik, and N. Stojanovic. User-driven ontology evolution management. In *Proceedings of the International Conference on Knowledge Engineering and Knowledge Management (EKAW 2002)*, pages 133–140, 2002.
33. E. Sunagawa, K. Kozaki, Y. Kitamura, and R. Mizoguchi. An environment for distributed ontology development based on dependency management. In *Proceedings of the 2nd International Semantic Web Conference (ISWC 2003)*, pages 453–468. Springer–Verlag, 2003.
34. O. Suominen and E. Hyvönen. Improving the quality of SKOS vocabularies with Skosify. In *Proceedings of the 18th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2012)*, pages 383–397. Springer-Verlag, October 2012.
35. O. Suominen, S. Pessala, J. Tuominen, M. Lappalainen, S. Nykyri, H. Ylikotila, M. Frosterus, and E. Hyvönen. Deploying national ontology services: From ONKI to Finto. In *Proceedings of the Industry Track at the International Semantic Web Conference 2014*. CEUR Workshop Proceedings, `http://CEUR-WS.org`, October 2014.
36. J. Tuominen, M. Frosterus, K. Viljanen, and E. Hyvönen. ONKI SKOS server for publishing and utilizing SKOS vocabularies and ontologies as services. In *Proceedings of the 6th European Semantic Web Conference (ESWC 2009)*, pages 768–780. Springer–Verlag, 2009.
37. K. Viljanen, J. Tuominen, and E. Hyvönen. Ontology libraries for production use: The Finnish ontology library service ONKI. In *Proceedings of the ESWC 2009, Heraklion, Greece*, pages 781–795. Springer–Verlag, 2009.
38. X. Wang, J. S. Almeida, and A. L. Oliveira. Ontology design principles and normalization techniques in the web. In *Proceedings of the 5th International Workshop on Data Integration in the Life Sciences (DILS 2008)*, pages 28–43. Springer–Verlag, June 2008.

Publication V

Kim Viljanen, Jouni Tuominen, Eetu Mäkelä and Eero Hyvönen. Normalized Access to Ontology Repositories. In *ICSC 2012: 2012 IEEE Sixth International Conference on Semantic Computing*, Palermo, Italy, 19-21 September 2012, pages 109–116, ISBN 978-1-4673-4433-3, IEEE, September 2012.

Normalized Access to Ontology Repositories

Kim Viljanen, Jouni Tuominen, Eetu Mäkelä and Eero Hyvönen
Semantic Computing Research Group (SeCo)
Aalto University, School of Science, and University of Helsinki
http://www.seco.tkk.fi/, firstname.lastname@aalto.fi

***Abstract*—Ontology repositories, such as NCBO Bioportal, ONKI and Cupboard, help finding and using ontologies on the Semantic Web. However, currently each ontology repository constitutes a separate island with its own user interface, APIs, users, ontology languages and set of ontologies. Because there is not a universal way to access all ontology repositories, doing global search, browsing, and inference over all available ontology repositories turns out to be technically difficult and is generally not done. Ontologies are not reused as much as they could and hence the full potential of ontologies is not achieved. To address the problem, we propose the Normalized Ontology Repository (NOR) approach to make the ontology repositories universally accessible while maintaining their unique functionalities and strengths. The SKOS language is used as the lowest common denominator for presenting the ontologies. In addition, a simple API for searching and accessing the ontologies is defined. As a proof-of-concept evaluation, we present three case implementations to demonstrate the NOR approach: 1) the distributed architecture of the ONKI repository, 2) the metasearch for ONKI and NCBO Bioportal, and 3) publishing informal ontological concept collections as NOR end-points, demonstrated with the semantic portal CultureSampo and the metadata editor SAHA.**

I. Introduction

Ontologies and ontology repositories have been considered to be a key resource for enabling the vision of the Semantic Web [1]–[3]. Ontology repositories are used for publishing, sharing and reusing ontologies and vocabularies for content indexing, information retrieval, content integration, and other purposes. Current implementations of ontology repositories include for example the NCBO BioPortal[1] [4], the Finnish Ontology Library Service ONKI[2] [5], Cupboard [6], the forthcoming Open Ontology Repository (OOR)[3] [2], and there are many other systems, too [3].

An ontology is a shared specification of a conceptualization, defining concepts of a specific area of interest to allow sharing of knowledge [7]. Ontologies typically contain textual information about the concepts, relations between the concepts and, perhaps most importantly, define the unique identifiers (the URI in the context of the Semantic Web) of the concepts. With the help of the identifiers, the concepts can be referred to, for example, as values in metadata. Therefore, one typical use case for ontology repositories is to support the user in finding relevant ontological concepts from the underlying ontologies. With the help of concept search and browsing functionalities, the user can find the best matching concepts for her needs. For example, if the user is creating metadata about an article about fishes, she could use an ontology repository containing an ontology about fishes to find out the correct URI for the concept "fish", and then use this identifier as the value in her metadata.

In addition to formal ontologies, a vast amount of other kinds of concept collections of various degrees of formality exist that could be useful as identifiers for the Semantic Web. We call them informal ontologies. One example of such an informal ontology is Wikipedia, where the URL of each Wikipedia page correspond to a concept. For example the BBC is using the Wikipedia identifiers for interlinking content [8] which is based on RDF representation of the Wikipedia, the DBpedia [9]. Other examples of informal ontologies include registries maintained by libraries, such as books and people (e.g. ULAN[4]), and identifiers maintained by other organizations, such as locations (e.g. GeoNames[5]). In addition, many websites and their underlying content management systems use site specific categories and other types of concept collections that could be useful for others, too. For example, both ebay[6] and Amazon[7] have extensive product categorizations.

A problem of current ontology repositories is that each system constitutes a separate island with its own functionalities and its own set of ontologies [10]. Each system has its own user-interface, own API, and support different ontology languages. This limits the user from using efficiently different ontology repositories together because, for example, searching for a specific concept simultaneously from many repositories is not possible but requires visiting each repository separately. As a solution to the problem, we propose a universal access method to ontology repositories based on a normalized presentation of the ontology content and a shared API. We call this the Normalized Ontology Repository (NOR) approach. In addition to ontology repositories, we argue that it is relevant to consider non-ontological concept collections also as valuable sources for concept identifiers. Therefore, we suggest that the NOR approach could be used for publishing such non-ontological sources, too.

In the following, we first discuss why there is a need for a multitude of ontology repositories with different functionalities instead of just creating one application or web service to address all the needs. Then we present the NOR approach

[1]http://bioportal.bioontology.org/
[2]http://onki.fi
[3]http://ontolog.cim3.net/cgi-bin/wiki.pl?OpenOntologyRepository
[4]http://www.getty.edu/research/tools/vocabularies/ulan/
[5]http://www.geonames.org
[6]http://www.ebay.com
[7]http://www.amazon.com

for creating universal access to different ontology repositories. After this, three implementations of the NOR approach are described. Finally, related work is presented, the results of this paper are summarized and discussed.

II. Motivation

To motivate our work, we discuss the following questions: Is there a need for different ontology repositories? Would simultaneous access to ontology repositories be useful? Do existing Semantic Web technologies and practices address the needs of ontology repositories?

A. One Size Does Not Fit All

Instead of using a multitude of different implementations for ontology repositories, one could argue that a single implementation could address all the different needs of the users, and the needs created by the different types of ontologies and application domains. In addition, there could be a single global ontology repository that would contain all ontologies. While both scenarios could theoretically be possible, we argue that there are a multitude of challenges in such approach due to the following reasons:

Different ontologies and user needs require different functionalities. For example, the ONKI ontology repository supports different types of ontologies and implements different visualizations, such as a geographical map interface for geographical ontologies and a tree visualization for concept hierarchies, as depicted in Fig. 1. This is done to address the different needs of different ontologies such as general (abstract) ontologies versus geographical ontologies. For example, the BioPortal has been designed originally to address the needs of the biomedical domain. Also different ontology languages require different technical implementations to maximize the benefits of the given formalism.

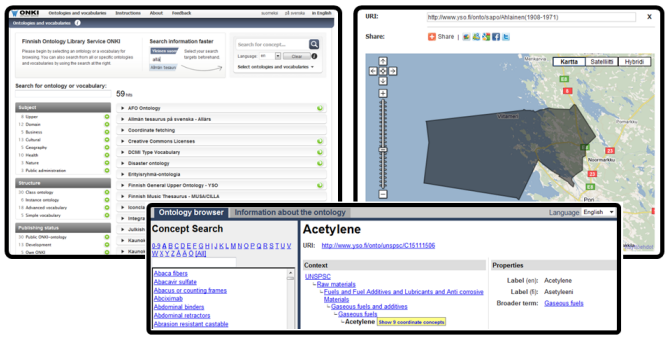


Fig. 1. Some of the user interfaces of the ONKI repository, including ontology listing, map visualization for geographical ontologies, and concept hierarchy visualization.

Some ontologies are not available as files but only as services. Typically, ontologies are published as files that can be uploaded to ontology repositories. For example, if the ontology changes constantly or if the size of the ontology is substantial, publishing the content as a file may not be practical. In addition, due to business reasons, some ontologies are not published as files but only as a service. In these cases, the ontology is only available via the specific API or a user-interface, but can not be uploaded to a shared ontology repository.

Security or business reasons may require using internal ontology repositories. For example, security reasons of an organization may require that selected ontologies are available only for internal use or that the server logs of who checked what ontological concepts remain confidential. Such requirements can be addressed with private, internal ontology repositories that are fully controlled by the organization itself.

All concept collections are not ontology repositories. As we pointed out in the introduction, a vast amount of informal ontological concept collections exists that are useful as identifiers for the Semantic Web, such as the Wikipedia, the various registries maintained by libraries and categorizations located in various websites, such as ebay and Amazon. A single ontology repository system most probably will not replace all of the different systems that are used for maintaining such informal ontologies and vocabularies of various degrees of formality.

B. Simultaneous Access to Repositories

The typical way to use ontology repositories and ontologies in applications is that the application developers choose a specific ontology repository and specific ontologies in advance which are then used in the application. The end-user does not have to bother what ontology repository or ontology to choose when using the application. We argue that the practice of using only one repository leads up to the following problems.

The optimal concept or ontology might not be found. Because searching simultaneously ontology repositories is not possible, the end-user or the application developer might not notice that there exists a suitable ontology or ontological concept for one's needs. Due to this, the quality of ontology-enabled applications and metadata may decrease when the perfect concepts are not found and used. Also redundant new ontologies and concepts may created if ontology developers do not find existing ontologies, which makes it more difficult to maintain the interoperability of data, because data originating from different sources is described using a different (redundant) ontology, which means that the ontologies must first be interlinked before they can be used for interlinking the underlying data.

High quality ontologies might be underused. The more a specific ontology is used, the more established it is to act as the de facto standard for representing concepts and metadata of its specific domain. This also increases the ontology developers and publishers benefits for creating and maintaining the ontology. If a specific ontology repository is not found by the potential users, then also the high quality ontologies contained in the repository will be underused, and the benefits from creating the ontology decreases. In addition, as pointed out before, informal ontological concept collections might not be considered as valuable sources for concept identifiers such as the Wikipedia or subject heading thesauri maintained

by libraries because they are not published in an existing ontology repository or made available using standard Semantic Web formats. Publishing them as ontology repositories would increase the benefits of existing work.

Ontologies are not interlinked between repositories. The ontology repositories are currently not acting as model citizens of the Semantic Web, since their ontological content is not interlinked between repositories. That is, the ontologies and the ontology repositories do not implement and follow the Linked Data [11] practices at the moment. For example, (automatic) linking to relevant concepts in other ontologies could help the users to find the best ontologies and concepts for each need.

Publishing same ontologies in many repositories creates challenges. Some ontologies may be published in many Ontology Repositories because they have been considered to be useful by the repository publisher, or because they have been uploaded by the ontology developer to as many repositories as possible to reach as many users as possible. Maintaining the same ontology in many repositories may lead to redundant work because new versions of the ontology has to be updated in all repositories. Also, some repositories may contain older versions of the ontology which creates compatibly problems when using metadata based on the ontologies.

Internal ontology repositories require maintenance. Internal, potentially confidential, ontology repositories may contain public ontologies. Maintaining the public ontologies to the latest versions require additional work. To avoid this, simultaneously access to both internal and public ontology repositories would be beneficial.

C. Shortcomings of Existing Technologies

General Semantic Web search engines, such as Swoogle[8] [12] and Sindice[9] [13] are not focused on ontologies but provide general search of all kind of RDF data. Ontology search engines, such as Falcons[10] [14] and Ontosearch2 [15] address ontology specific needs, but do not address the problem of accessing informal ontology repositories. In addition, all of the previously mentioned search engines are based on crawling the ontology sources, which means, that they are may not always be up-to-date. In addition, ontologies that are only available as services, via an API or user-interface, may not be indexed.

Ontologies are represented using various languages, such as the Semantic Web languages RDFS, OWL and SKOS, Common Logic, Excel, HTML, database tables, and application specific languages. A shared practice is missing on how to publish ontologies on the Semantic Web [2].

SPARQL[11] is the standard way to provide an application interface to Semantic Web databases, and it can be used also to access ontology repositories. However, implementing a SPARQL end-point can be difficult if the underlying system is not based on Semantic Web technologies. Making SPARQL queries require also advance knowledge on what ontology language has been used to be able to make a matching query and to interpret the result.

To conclude our analysis, in the foreseeable future there will be many different ontology repositories. Accessing simultaneously these repositories would be useful and is not solved in an optimal way with current technologies.

III. The Normalized Ontology Repositories

As a solution to the problems presented above, we propose the Normalized Ontology Repository (NOR) approach. NOR consist of 1) a normalized presentation for ontology concepts, making thus the different ontology language schemas interoperable, and 2) a simple API for accessing the ontology repository.

A. Normalized Representation of Ontological Concepts

Ontologies are presented using different ontology languages, such as OWL, RDFS and SKOS, and there exists many informal ontologies, too. From the interoperability point of view, this creates a problem, because each ontology language must handled as a separate case. In the worst case, an application developer have to handle ontologies presented in many different languages to build an application that utilizes ontologies. Due to this, for example, the ONKI repository has a rule-based configuration language to adjust the system to support ontologies represented in various kinds of RDF based languages.

To avoid complicated mappings and inference of hierarchical and other relations, we propose that each ontology repository should provide a normalized, dumbed down presentation of the ontology concepts in addition to the native format of the ontology. As the normalization language we suggest using the RDF based Simple Knowledge Organization System (SKOS)[12], which is a RDF based language for presenting thesauri, classification schemes, subject heading systems and taxonomies within the framework of the Semantic Web. SKOS is by design intended to serve as a common denominator between different modeling approaches and therefore we decided to use it compared to other alternatives, such as OWL or RDFS.

Hiding ontological details makes it easier for the applications using the NOR compatible ontology repositories. After finding an interesting concept, the user can be directed to the specific Ontology Repository with its full functionality for using the specific ontology. Our intention is to make it easier to access the basic information of ontological concepts in an unified way, not to restrict the user from using the original, full-blown ontology languages and functionalities of the underlying ontology repositories for specific needs.

In practice, a NOR compatible ontology repository must provide a concept lookup method:

- *concept?uri=[concept identifier]*

[8] http://swoogle.umbc.edu
[9] http://sindice.com
[10] http://ws.nju.edu.cn/falcons
[11] http://www.w3.org/TR/rdf-sparql-query/
[12] http://www.w3.org/2004/02/skos/

which returns the normalized SKOS version of the given concept, identified by the concept URI. For example, to get the normalized concept representation of *yso:p907* from the ONKI ontology repository, the lookup request URL is:

```
http://onki.fi/nor/concept?uri=http%3A%2F%2F
www.yso.fi%2Fonto%2Fyso%2Fp907
```

which returns the following SKOS representation[13] of the given concept followed by the (optional) native representation:

```
# Namespace declarations omitted
# Normalized SKOS representation begins
<http://onki.fi/nor/concept?uri=http%3A%2F%2F
www.yso.fi%2Fonto%2Fyso%2Fp907>
   a skos:Concept;
   skos:prefLabel "fish"@en, "kala"@fi;
   skos:broader
     <http://onki.fi/nor/concept?uri=http%3A
%2F%2Fwww.yso.fi%2Fonto%2Fyso%2Fp6580>;
#additional properties omitted
#link to the native concept format
   nor:describes yso:p907
.
# Native representation begins (optional)
yso:p907
   a ysometa:Concept;
   ysometa:prefLabel "fish"@en, "kala"@fi;
   #...additional properties omitted
.
```

The SKOS presentation above describes key information about the given concept (*yso:p907*) such as the labels (in English "fish", in Finnish "kala"), and the URL to the normalized broader concept *yso:p6580* (foods). In addition, the native representation *yso:p907* is also presented as part of the normalized concept lookup response.

To avoid cluttering the native presentation by adding additional RDF triplets to it, the native and normalized formats are kept apart from each other with the following RDF property[14]:

- *nor:describes*

The property is used for referring to the native concept presentation from the normalized SKOS representation. To avoid making unintended conclusions, we did not use, for example, the *owl:sameAs* property which would have meant that the normalized and the native presentations would refer to the same thing, which may not be true.

Finally, in some cases the ontology repository publisher may have decided to use SKOS as the native representation for the concepts. If so, the *nor:describes* relation and the native representation can be omitted.

B. Concept Search

To make searches to a NOR compatible ontology repository we define the following method:

- *search?q=[query]&l=[language]*

The search method is used for finding concepts matching the given query string and language. Currently, the query string can only contain a keyword, but in future the query language may be extended. The method returns a list of matching concepts presented using a JSON based response format. Other result languages and formats may be considered in the future, but we deemed this representation to be simpler than, for example, representing the same information as an ordered list in RDF.

For example, a search for "fish" to the ONKI ontology repository is done with the following URL:

```
http://onki.fi/nor/search?q=fish&l=en
```

The system responds with the following result:

```
{"results" : [
  {"concept-label" : "fish",
   "concept-label-language" : "en",
   "concept" :
       "http://www.yso.fi/onto/yso/p907",
   "normalized-concept" :
       "http://onki.fi/nor/concept?uri=http%3A
       %2F%2Fwww.yso.fi%2Fonto%2Fyso%2Fp907",
   "native-concept" :
       "http://www.yso.fi/onto/yso/p907",
   "ontology-abbreviation" : "yso",
   "ontology-label" : "Finnish General Upper
       Ontology",
   "ontology-label-language" : "en",
   "ontology-uri" :
       "http://www.yso.fi/onto/yso"
  }
  ...
 ],
 "metadata" : {"containingHitsAmount" : 50,
               "moreHitsAmount" : 1467}
}
```

In the result, *concept* is the URI of the concept, *normalized-concept* is the URL of the normalized representation of the given concept, and *native-concept* is the URL to the native representation of the concept.

C. Ontology Repository Metadata

To find NOR compatible ontology repositories, a list of repositories that conform to the NOR principles would be helpful. However, to avoid the problems of centralized systems, we do not require ontology repositories to publish information about themselves to any specific registries.

To help finding suitable repositories and ontologies for one's need, we suggest that the NOR ontology repositories publish metadata about the available ontologies using the following method:

- *ontologies*

which returns metadata about the ontologies in the repository and the NOR end-point URL of each ontology. The metadata of the ontologies can be represented using, for example, the Vocabulary of Interlinked Datasets (voiD)[15] metadata language. Additional information about the ontology, such as the title and description, may be expressed using e.g. the Dublin Core metadata schema, the Ontology Metadata Vocabulary (OMV)[16], and the upcoming Catalogue Vocabulary (dcat)[17].

[13] presented using the RDF Turtle syntax

[14] nor namespace: http://purl.org/finnonto/schema/nor

[15] http://rdfs.org/ns/void#

[16] http://omv2.sourceforge.net

[17] http://www.w3.org/egov/wiki/Data_Catalog_Vocabulary

To express the URL of the NOR end-point for a given ontology, we define the following RDF property:

- *nor:endpoint*

For example, in the case of the ONKI ontology repository, the ontology metadata is available at:

```
http://onki.fi/nor/ontologies
```

which returns following metadata (excerpt):

```
# Namespace declarations omitted
<http://www.yso.fi/onto/yso>
  a void:Dataset ;
  dc:title "Finnish General
   Upper Ontology"@en ;
  dc:creator
   <http://www.yso.fi/onki-ns/onki/Finnonto> ;
  dc:license <http://creativecommons.org/
licenses/by/3.0/> ;
  foaf:homepage
   <http://www.seco.tkk.fi/ontologies/yso>,
  nor:endpoint <http://onki.fi/nor/> .
...
```

The metadata can be used for creating for example a catalogue of NOR compatible ontology repositories and concept collections. In addition, this could be used for implementing a metasearch service to search simultaneously multiple underlying NOR compatible ontology repositories.

D. General Remarks on the API

We argue, that a simple HTTP API is easy to implement both for the ontology repository developers and for the developers that want to access the NOR compatible ontology repositories. In addition, a simple API is easy to implement even if the underlying ontology repository is not based on RDF but is, for example, a relational database of people, which could be highly relevant to publish as a NOR endpoint. Thus, compared to e.g. using the RDF query language SPARQL, the simple API approach makes it easier for both publishers and users to benefit of the NOR network.

This does not limit, however, the underlying ontology repositories from implementing in addition, for example, a SPARQL end-point. A key idea behind NOR is that the native functionalities of the underlying ontology repositories are available for users that need more functionalities than what the simple NOR API and normalized presentation can provide.

The proposed API is summarized in the Table I.

TABLE I
THE NORMALIZED ONTOLOGY REPOSITORY API.

Method name	Parameters	Return value
concept	concept identifier	normalized concept representation
search	query string, language	matching concepts
ontologies	-	ontology metadata

IV. CASE STUDIES AND EVALUATION

To analyze the idea of NOR, we have implemented three proof-of-concept prototypes which will be presented and discussed in the following.

The NOR approach generalizes and unifies experiences gained from our work on the ONKI repository and the ONKI API. Therefore, a majority of the following case studies are based on ONKI, viewed from the NOR perspective. The functionalities of the NOR API is a subset of the ONKI API's functionalities, however, a key difference is that the ONKI API represents the SKOS description of a concept in an ONKI specific JSON format to avoid the overhead of parsing RDF in the ONKI frontend, but in the NOR API we propose using RDF to represent this information.

A. NOR as an Internal Architecture

The ONKI SKOS ontology server [16] has been used for publishing over 70 ontologies in the Finnish Ontology Library Service ONKI [5], which has been running as a pilot service from September 2008. The system is in living lab use with ca. 10 000 unique human visitors monthly[18], and there are over 300 registered users of the APIs and widgets. Even though ONKI SKOS supports especially vocabularies presented in SKOS, the server can be used for publishing ontologies presented in RDF(S) and restricted OWL. To access the multitude of ONKI SKOS servers, the ONKI system implements a front-end service for making metasearches to the ONKI SKOS and other ONKI back-ends using a shared HTTP API (see Fig. 2). The back-ends and their respective ontologies are described with metadata to enable the front-end to locate the available ontologies and to display information about the ontologies to the users.

Searching for concepts using an ONKI backend server is done with its HTTP API method *search*, which returns concepts matching to the query string in a specified language. The *getFullPresentation* method returns all information about a given concept, such as the preferred and alternative labels, the transitive parent concept tree, and the related concepts. Independently of the language each ontology is presented in, each concept is always returned in a uniform SKOS inspired JSON format which describe the normalized basic information of the given concept.

Building on this underlying distributed architecture, three clients have been designed and implemented. The ONKI3 Browser[19] is a metasearch and browsing user interface for accessing the ONKI SKOS and other back-end servers. For example, making a global query to all ontology servers can be done. Also, a directory listing of the ontologies in the ONKI Ontology Repository is provided based on the metadata about the published ontologies. The ONKI3 user interface was mostly implemented using PHP[20].

Another client is the JavaScript-based ONKI Selector widget [17] for adding ontological concept search to HTML forms. The third client is a simple URI resolver for dereferencing the end-user's ontology concept URI requests to a suitable representation provided via the ontology repository network, such as HTML or RDF.

[18]Measured with Google Analytics.
[19]http://onki.fi/en/browser
[20]http://www.php.net/

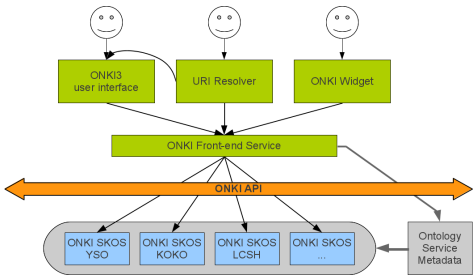


Fig. 2. The ONKI architecture is based on a distributed metasearch approach.

The loosely coupled ONKI architecture has turned out to be a flexible and modularized approach for implementing an ontology repository consisting of multiple back-end ontology servers. The normalized representation of the underlying ontology repositories have made it easy to implement a user-interface for accessing all underlying repositories. Making multiple HTTP requests to back-end servers may be slow in the worst case, but in our test implementation this lag has not been a problem.

B. Searching Simultaneously BioPortal and ONKI

To test the NOR approach in a distributed setting of multiple independent ontology repositories, we implemented a proof-of-concept metasearch prototype to search simultaneously the ONKI SKOS [16] servers described above and the NCBO BioPortal [4]. The NCBO BioPortal is an open repository of biomedical ontologies and it has been used for publishing over 200 ontologies [4]. BioPortal provides functionalities, such as concept and ontology search and browsing, peer reviewing of the ontologies, and support for creating and viewing mappings between ontologies.

The ONKI-BioPortal metasearch prototype allows the user to find the relevant concepts from the participating ontology repositories, without having to know in advance which repository to make the search to.

Since the ONKI front-end [5] was already designed using a metasearch approach, the ONKI-BioPortal prototype was implemented by creating a wrapper for BioPortal which implements the ONKI API's search and concept lookup (*getFullPresentation*) methods. When calling the wrapper, it makes requests to BioPortal, parses BioPortal's XML messages, and transforms them to the ONKI JSON format. Since the BioPortal API does not contain a concept lookup method that would return all information about the specific concept with a single request, multiple HTTP REST requests have to be made to get all the needed information about a concept.

Fig. 3 presents the ONKI-BioPortal search prototype user-interface displaying the result for a metasearch query for "fish product". The result consists of 22 hits which are found from the BioPortal and the ONKI SKOS back-ends. The hits originating from BioPortal are labeled as "BioPortal" for demonstrating purposes, but in actual use, instead of "BioPortal" the name of the ontology should be displayed.

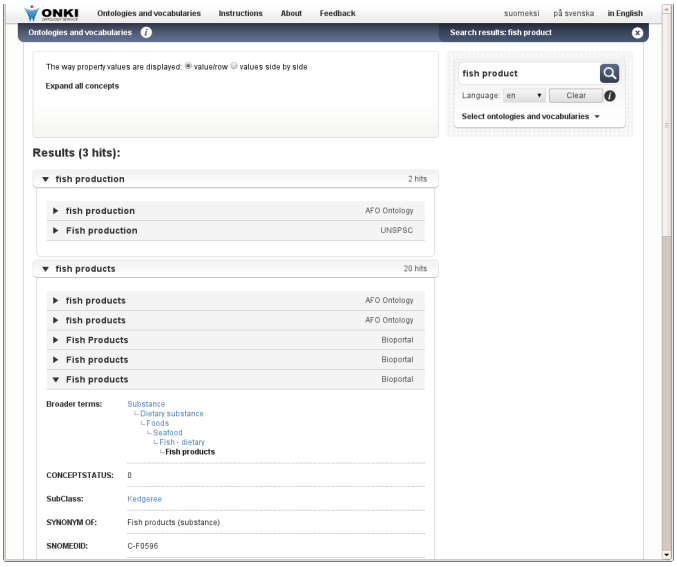


Fig. 3. The user-interface of the ONKI-BioPortal metasearch prototype.

C. NOR for Informal Ontological Concept Collections

Besides Ontology Repositories, applications often need to refer also to informal ontological concept collections, such as authority or place databases. However, the functionalities required for such data sources are usually very similar to those required for ontology repositories. For example, in an editor environment, similar semantic autocompletion search functionalities are used for both ontological and non-ontological concept collections, along with the same functionality for describing and visualizing the possible choices returned from such a search. Informal ontological concept collections also often change more rapidly than their ontological counterparts, so it makes even more sense to access the original system through programmatic APIs than exporting and publishing the data in a ontology repository. In order to test how the NOR approach fared in the context of such informal ontological concept collections, the ONKI API was implemented in two applications: the semantic portal CultureSampo [18] and the SAHA metadata editor [19]. Both are Semantic Web applications, but their focus is not on ontologies but to display and edit all kinds of semantic data.

For CultureSampo, the ONKI API was actually implemented to benefit those using SAHA to edit data. This was because the CultureSampo database contains, for example, a large number of places, people and organizations that are useful to people indexing new content. For added freedom, the CultureSampo ONKI API was parametrized, so that the types of objects that search operations return can be specified dynamically. This way, one can say for example that they want an autocompletion facility of all the organizations, all the places, or all the historical events in CultureSampo.

While SAHA was already a client to the ONKI API of the ONKI Ontology Repository and CultureSampo, the API was

implemented also into SAHA itself. This was done to make possible the creation of a network of dynamically updated, collaboratively curated concept collections. The multiple projects using the SAHA editor to index content often need to add new places, organizations or people to their list of reference values. However, until now, these have all resided in the private data spaces of the different projects using SAHA. Now, the intention is to move these created concepts into SAHA projects of their own, so that one SAHA project will hold collaboratively curated place database, while another contains a database of organizations and people. These can then be linked through the ONKI APIs to each other, as well as to the primary indexing projects. In this way, the various projects can start to directly benefit each other.

V. Related Work

The work is partially based on our previous work on the national ontology library ONKI [5], [20] and is related to the open ontology repository (OOR)[21] initiative which aims at developing an interoperability infrastructure for ontologies [2].

Compared to more general methods of accessing RDF data, such as SPARQL[22] and Linked Data [11], the NOR approach focuses on ontologies. For example, when searching for concepts with the NOR API, one does not need to know what RDF properties are used in the data to express the labels. In addition, the ontology repositories can be optimized to respond quickly to specific API queries. A normalized presentation of ontological concepts (SKOS) could, however, also be beneficial for querying the data via SPARQL, and browsing the ontology repositories as linked data. For example, one does not have to know which specific hierarchical relation (e.g. *rdfs:subClassOf* or *skos:broader*) has been used, because the normalized hierarchical relation is constant.

APIs for accessing ontologies and vocabularies published by other authors previously include the SKOS API[23] and the OWL API[24]. Compared to them, the NOR approach provides a higher abstraction, independent from specific ontology languages, and a lightweight and simple API. Compared to the APIs of BioPortal [4], Swoogle[25] [12], Watson[26] [21], ONKI SKOS and others, the NOR API focuses on a few basic methods that reflects the basic functionality of ontology repositories, e.g. concept search.

The Ontosearch2 [15] does a automatic complexity reduction of ontologies to ensure answering the ontology search queries within a specific time limit. This automatic approach however require using the OWL ontology language which is a limitation since many ontologies are not presented in that specific language. In contrast, the NOR approach is based on defining the normalized language and the simple API with the goal of publishing both ontology repositories and informal ontological concept collections as uniform services.

Ontology Repositories such as BioPortal and Cupboard support publishing interlinked ontologies, but the ontologies have to be uploaded into a centralized service for a global search. On the other hand, the OOR [2] initiative intends to design an Ontology Repository framework that addresses the needs of all users, and includes an inter-repository content change protocol to keep the different OOR repositories up to date. In contrast to these, the NOR approach does not restrict the ontology publishers in where to publish the ontologies or what software to use. Instead, the ontologies can be published using an ontology service that is optimized for the specific ontology and the user's needs. If the organization wants to promote and make their ontologies available to the NOR users, they can implement the NOR API to make their repository compatible with other NOR repositories. If needed, the NOR API of a repository can be restricted to selected users or made publicly available for anybody.

Compared to the Open Knowledge Base Connectivity (OKBC) specification[27] and the agent communications languages FIPA-ACL[28] and KQML[29], which all can also be used to access ontological information, the NOR approach is more focused on the specific use-cases of finding ontologies and ontology concepts, and to get relevant information about them.

The OntoCAT is a programming interface to query multiple ontology repositories seamlessly from an application [10]. A wrapper is implemented for each supported ontology repository, such as the NCBO BioPortal. In comparison, to avoid wrappers, the NOR approach is based on defining a shared, unified representation for the ontology repositories. A well-known limitation of wrappers is that changes in the underlying representation often breaks the wrapper.

VI. Discussion

This paper argues that ontology repositories should be made accessible using a shared API that would provide a simple but universal methods for accessing the repositories in a uniform way. In addition, the ontologies should be presented using a normalized concept representation.

The NOR approach has been evaluated with three case studies: The ONKI ontology repository case study demonstrates using the NOR approach for building an ontology service consisting of over 70 underlying back-ends with over 10 000 unique monthly users. The NCBO BioPortal and ONKI case study demonstrates using the NOR approach for creating a global search and browsing user-interface for accessing independent distributed ontology repositories. Finally, the SAHA metadata editor and the CultureSampo semantic portal case study demonstrates that the NOR approach can be used for accessing non-ontological concept collections.

The outcome of this work is that the NOR approach is feasible for providing a unified access to a multitude of

[21] http://ontolog.cim3.net/cgi-bin/wiki.pl?OpenOntologyRepository
[22] http://www.w3.org/TR/rdf-sparql-query/
[23] http://www.w3.org/2001/sw/Europe/reports/thes/skosapi.html
[24] http://owlapi.sourceforge.net/
[25] http://swoogle.umbc.edu/
[26] http://watson.kmi.open.ac.uk/
[27] http://www.ai.sri.com/ okbc/spec.html
[28] http://www.fipa.org/repository/aclspecs.html
[29] http://www.cs.umbc.edu/csee/research/kqml/

ontology repositories. This makes it possible to provide for example global search and global browsing functionalities to a collection of separate underlying ontology repositories. At the same time, the NOR does not restrict the individual ontology repository providers from creating advanced ontology, business, and user specific implementations because the relation between the normalized representation and the native representation is kept intact.

The NOR approach allows the ontology user to find relevant concepts and ontology repositories in cases where the correct ontology repository is not known in advance or when many ontology repositories are used simultaneously. After finding the relevant repository, the user may access the underlying ontology repository for full-blown functionalities. For organizations that maintain an internal ontology repository, the NOR approach makes it possible to make simultaneous queries to repositories outside the organization. For the ontology publishers, implementing the NOR API increases the findability of the ontologies and therefore the benefits of publishing the ontology in the first place.

Future work includes developing further the API and its methods to support, for example, restricting queries to a specific ontology, specific subpart of the ontology or to a specific concept type. The normalized concept representations could be improved by introducing links between ontologies in the spirit of Linked Data. Such mappings between ontologies could be produced potentially automatically by creating a matching application on the top of the NOR compatible ontology repositories. NOR based metasearch would benefit from a ranking algorithm for ordering the results originating from different underlying ontology repositories. Finally, to evaluate the full potential of the approach, formal and informal ontology repositories should implement the NOR API.

Acknowledgements: This work is part of the National Semantic Web Ontology project in Finland[30] (FinnONTO, 2003-2012), funded mainly by the National Technology and Innovation Agency (Tekes) and a consortium of 38 organizations. We thank Osma Suominen, the Semantic Computing Research Group, and the OOR network for fruitful discussions.

REFERENCES

[1] E. Hyvönen, K. Viljanen, J. Tuominen, and K. Seppälä, "Building a national semantic web ontology and ontology service infrastructure—the FinnONTO approach," in *Proceedings of the ESWC 2008, Tenerife, Spain*. Springer–Verlag, 2008.

[2] K. Baclawski and T. Schneider, "The open ontology repository initiative: Requirements and research challenges," in *Proceedings of Workshop on Collaborative Construction, Management and Linking of Structured Knowledge at the ISWC 2009*, Washington DC., USA, October 2009.

[3] M. d'Aquin and N. F. Noy, "Where to publish and find ontologies? A survey of ontology libraries," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 11, pp. 96–111, Mar. 2012. [Online]. Available: http://dx.doi.org/10.1016/j.websem.2011.08.005

[4] N. F. Noy, N. H. Shah, P. L. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D. L. Rubin, M.-A. Storey, C. G. Chute, and M. A. Musen, "BioPortal: ontologies and integrated data resources at the click of a mouse," *Nucleic Acids Research*, vol. 37, no. Web Server issue, pp. 170–173, 2009.

[5] K. Viljanen, J. Tuominen, and E. Hyvönen, "Ontology libraries for production use: The Finnish ontology library service ONKI," in *Proceedings of the ESWC 2009, Heraklion, Greece*. Springer–Verlag, 2009.

[6] M. d'Aquin and H. Lewen, "Cupboard - a place to expose your ontologies to applications and the community," in *Proceedings of the ESWC 2009*. Heraklion, Greece: Springer–Verlag, June 2009, pp. 913–918.

[7] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowledge Acquisition*, vol. 5, no. 2, pp. 199–220, 1993.

[8] G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer, and R. Lee, "Media meets semantic web – how the bbc uses dbpedia and linked data to make connections," in *Proceedings of the ESWC 2009, Heraklion, Greece*. Springer–Verlag, 2009.

[9] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, "Dbpedia a crystallization point for the web of data," *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, no. 7, p. 154165, 2009.

[10] T. Adamusiak, T. Burdett, N. Kurbatova, K. J. van der Velde, N. Abeygunawardena, D. Antonakaki, M. Kapushesky, H. Parkinson, and M. Swertz, "Ontocat – simple ontology search and integration in java, r and rest/javascript," *BMC Bioinformatics*, vol. 12, no. 1, p. 218, 2011. [Online]. Available: http://www.biomedcentral.com/1471-2105/12/218

[11] C. Bizer, R. Cyganiak, and T. Heath, "How to publish linked data on the web," http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/, July 27 2007.

[12] L. Ding, T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. C. Doshi, and J. Sachs, "Swoogle: A Search and Metadata Engine for the Semantic Web," in *Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management*. ACM Press, November 2004.

[13] E. Oren, R. Delbru, M. Catasta, R. Cyganiak, H. Stenzhorn, and G. Tummarello, "Sindice.com: a document-oriented lookup index for open linked data." *IJMSO*, vol. 3, no. 1, pp. 37–52, 2008. [Online]. Available: http://dblp.uni-trier.de/db/journals/ijmso/ijmso3.html

[14] Y. Qu and G. Cheng, "Falcons concept search: A practical search engine for web ontologies," *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, vol. 41, no. 4, pp. 810–816, 2011.

[15] E. Thomas, J. Z. Pan, and D. Sleeman, "ONTOSEARCH2: Searching Ontologies Semantically," in *Proceedings of the OWLED 2007 Workshop on OWL: Experiences and Directions*, 2007, online publication: http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-258/.

[16] J. Tuominen, M. Frosterus, K. Viljanen, and E. Hyvönen, "ONKI SKOS server for publishing and utilizing SKOS vocabularies and ontologies as services," in *Proceedings of the ESWC 2009, Heraklion, Greece*. Springer–Verlag, 2009.

[17] K. Viljanen, J. Tuominen, and E. Hyvönen, "Publishing and using ontologies as mash-up services," in *Proceedings of the 4th Workshop on Scripting for the Semantic Web (SFSW2008), 5th European Semantic Web Conference 2008 (ESWC 2008)*, June 1-5 2008.

[18] E. Hyvönen, E. Mäkelä, T. Kauppinen, O. Alm, J. Kurki, T. Ruotsalo, K. Seppälä, J. Takala, K. Puputti, H. Kuittinen, K. Viljanen, J. Tuominen, T. Palonen, M. Frosterus, R. Sinkkilä, P. Paakkarinen, J. Laitio, and K. Nyberg, "CultureSampo – Finnish culture on the semantic web 2.0. Thematic perspectives for the end-user," in *Proceedings, Museums and the Web 2009, Indianapolis, USA*, April 15-18 2009.

[19] J. Kurki and E. Hyvönen, "Collaborative metadata editor integrated with ontology services and faceted portals," in *Workshop on Ontology Repositories and Editors for the Semantic Web (ORES 2010), the Extended Semantic Web Conference ESWC 2010, Heraklion, Greece*. CEUR Workshop Proceedings, http://ceur-ws.org/, June 2010.

[20] K. Viljanen, J. Tuominen, M. Salonoja, and E. Hyvönen, "Linked open ontology services," in *Workshop on Ontology Repositories and Editors for the Semantic Web (ORES 2010), the Extended Semantic Web Conference ESWC 2010*. CEUR Workshop Proceedings, http://ceur-ws.org/, June 2010.

[21] M. d'Aquin, M. Sabou, E. Motta, S. Angeletou, L. Gridinoc, V. Lopez, and F. Zablith, "What can be done with the semantic web? an overview of watson-based applications," in *5th Workshop on Semantic Web Applications and Perspectives, SWAP 2008*, 2008.

[30] http://www.seco.tkk.fi/projects/finnonto/

Publication VI

Jouni Tuominen, Nina Laurenne, and Eero Hyvönen. Biological Names and Taxonomies on the Semantic Web – Managing the Change in Scientific Conception. In *The Semantic Web: Research and Applications: 8th Extended Semantic Web Conference, ESWC 2011, Heraklion, Crete, Greece, May 29 – June 2, 2011, Proceedings, Part II*, Grigoris Antoniou, Marko Grobelnik, Elena Simperl, Bijan Parsia, Dimitris Plexousakis, Pieter De Leenheer, and Jeff Pan (editors), Lecture Notes in Computer Science, volume 6644, pages 255–269, ISBN 978-3-642-21063-1, Springer-Verlag, June 2011.

Biological Names and Taxonomies on the Semantic Web – Managing the Change in Scientific Conception

Jouni Tuominen, Nina Laurenne, and Eero Hyvönen

Semantic Computing Research Group (SeCo)
Aalto University School of Science and the University of Helsinki
firstname.lastname@aalto.fi
http://www.seco.tkk.fi/

Abstract. Biodiversity management requires the usage of heterogeneous biological information from multiple sources. Indexing, aggregating, and finding such information is based on names and taxonomic knowledge of organisms. However, taxonomies change in time due to new scientific findings, opinions of authorities, and changes in our conception about life forms. Furthermore, organism names and their meaning change in time, different authorities use different scientific names for the same taxon in different times, and various vernacular names are in use in different languages. This makes data integration and information retrieval difficult without detailed biological information. This paper introduces a meta-ontology for managing the names and taxonomies of organisms, and presents three applications for it: 1) publishing biological species lists as ontology services (ca. 20 taxonomies including more than 80,000 names), 2) collaborative management of the vernacular names of vascular plants (ca. 26,000 taxa), and 3) management of individual scientific name changes based on research results, covering a group of beetles. The applications are based on the databases of the Finnish Museum of Natural History and are used in a living lab environment on the web.

1 Introduction

Exploitation of natural resources, urbanisation, pollution, and climate changes accelerate the extinction of organisms on Earth which has raised a common concern about maintaining biodiversity. For this purpose, management of information about plants and animals is needed, a task requiring an efficient usage of heterogeneous, dynamic biological data from distributed sources, such as observational records, literature, and natural history collections. Central resources in biodiversity management are names and ontological taxonomies of organisms [1,19,20,3,4]. Animal ontologies are stereotypical examples in the semantic web text books, but in reality semantic web technologies have hardly been applied to managing the real life taxonomies of biological organisms and biodiversity on the web. This paper tries to fill this gap.[1]

[1] We discuss the taxonomies of contemporary species, not 'phylogenetic trees' that model evolutionary development of species, where humans are successors, e.g., of dinosaurs.

G. Antoniou et al. (Eds.): ESWC 2011, Part II, LNCS 6644, pp. 255–269, 2011.

Managing taxonomies of organisms provides new challenges to semantic web ontology research. Firstly, although we know that lions are carnivores, a subclass of mammals that eat other animals, the notion of 'species' in the general case is actually very hard to define precisely. For example, some authors discuss as many as 22 different definitions of the notion of species [16]. Secondly, taxonomic knowledge changes and increases due to new research results. The number of new organism names in biology increases by 25,000 every year as new taxa to science are discovered [11]. At the same time, the rate of changes in existing names has accelerated by the implementation of molecular methods suggesting new positions to organisms in taxonomies. Thirdly, biological names are not stable or reliable identifiers for organisms as they or their meaning change in time. Fourthly, the same name can be used by different authors to refer to different taxa (units of classification that commonly have a rank in the hierarchy), and a taxon can have more than one name without a consensus about the preferred one.

As a result, biological texts are written, content is indexed in databases, and information is searched for using different names and terms from different times and authorities. In biological research, scientific names are used instead of common names, but in many applications vernacular names in different languages are used instead. Data fusion is challenging and information retrieval without deep biological knowledge is difficult.

We argue that a shared system for publishing and managing the scientific and vernacular names and underlying conceptions of organisms and taxonomies is needed. From a research viewpoint, such a system is needed to index research results and to find out whether a potential new species is already known under some name. Biological information needed by environmental authorities cannot be properly indexed, found or aggregated unless the organism names and identifiers are available and can be aligned. For amateur scientists and the public, aligning vernacular names to scientific names and taxonomies is often a prerequisite for successful information retrieval.

This paper presents a meta-ontology and its applications addressing these problems. Our research hypothesis is that semantic web technologies are useful in practise in modelling change in the scientific perception of biological names and taxonomies, for creating a platform for collaboratively managing scientific knowledge about taxonomies, and for publishing taxonomies as ontology services for indexing and information retrieval purposes in legacy systems.

In the following, biological classification systems are first discussed and a meta-ontology TaxMeOn for defining such systems is presented [13]. Three use case applications of the meta-ontology are then discussed: a system for managing vascular plant names collaboratively (26,000 species) based on the SAHA metadata editor [12], application of the ONKI ontology service [25] for publishing taxonomic species lists on the semantic web (over 80,000 taxa of mammals, birds, butterflies, wasps, etc.), and a more focused application for managing the names and scientific findings of the Afro-tropical beetle family Eucnemidae. Finally, contributions of our work are summarised, related work discussed, and directions for further research are outlined.

2 Biological Names and Taxonomies

The scientific name system is based on the Linnean binomial name system where the basic unit is a species. Every species belongs to some genus and every genus belongs to a higher taxon. A scientific name often has a reference to the original publication where it was first published. For example, the scientific name of the bumblebee, *Apis mellifera* Linnaeus, 1758, means that Linnaeus published the description of the bumblebee in 1758 (in Systema Naturae 10th edition) and that bumblebee belongs to the genus *Apis*. The upper levels of the taxonomic hierarchy do not show in a scientific name. A confusing feature of scientific names is that the meaning of the name may change although the name remains the same. Taxon boundaries may vary according to different studies, and there may be multiple simultaneous views of taxon limits of the same organism group. For example, a genus may be delimited in three ways and according to each view different sets of species are included in the genus as illustrated in Fig. 1. These differing views are *taxonomic concepts*. The usage of the correct name is not enough, and Berendsohn [1] suggested that taxonomic concepts should be referred to by an abbreviation *sec* (*secundum*) after the authors name to indicate in which meaning the name is used.

The nature of a biological name system is a change, as there is no single interpretation of the evolution. Typically there is no agreement if the variation observed in an organism is taxon-specific or shared by more than one taxon, which makes the name system dynamic. For example, the fruit fly *Drosophila melanogaster* was shifted into the genus *Sophophora*, resulting in a new name combination *Sophonophora melanogaster* [7]. The most common taxonomic changes and their implications to the scientific names are the following: 1) A species has been shifted to another genus - the genus name changes. 2) One species turns out to be several species - new species are described and named, and the old name remains the same with a narrower interpretation. 3) Several species are found to be just one species - the oldest name is valid and the other names become its synonyms.

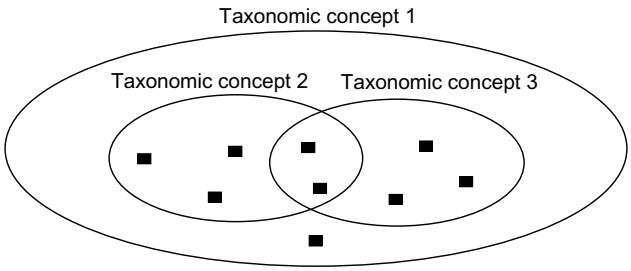


Fig. 1. A genus is delimited in three different ways according to three different studies. Black squares indicate species.

Species lists catalogue organisms occurring in a certain geographical area, which may vary from a small region to global. Often species lists contain valid taxon names with author information and synonyms of the valid names. They are snapshots of time and used especially by environmental authorities. The problem with species lists is that not all organism groups are catalogued and changes are not necessarily recorded in the species lists. Traditionally printed lists tend to be more detailed than online lists and their status is higher.

Species lists often follow different hierarchies and species may be associated with different genera according to the person who published the list. The hierarchy in a species list is a compromise that combines several studies, and the author can subjectively emphasise a view that he/she wishes. A taxon may also have different taxonomic ranks in literature, for example the same taxon can occur both as a species and a subspecies.

Common names tend to have regional variation and they do not indicate hierarchy unlike scientific names. Vernacular names have an important role in everyday language, but due to the variation and vagueness, they have little relevance in science. Vernacular names are used mainly in citizen science.

3 TaxMeOn – Meta-ontology of Biological Names

We have developed a meta-ontology for managing scientific and vernacular names. The ontology model consists of three parts that serve different purposes: 1) name collections, 2) species lists, and 3) name changes resulting from research. These parts are manageable separately, but associations between them are supported. Being a meta-ontology, TaxMeOn defines classes and properties that can be used to build ontologies. The ontologies can be used for creating semantic metadata for describing e.g. observational data or museum collections. TaxMeOn is based on RDF using some features of the OWL. The model contains 22 classes and 53 properties (61 including subproperties), of which ten classes and 15 properties are common to all the three parts of the model[2].

The core classes of TaxMeOn express a taxonomic concept, a scientific name, a taxonomic rank, a publication, an author, a vernacular name, and a status of a name. Taxonomic ranks are modelled as classes, and individual taxa are instances of them, for example the species forest fir *forrestii* (belongs to the genus *Abies*) is an instance of the class *Species*. The model contains 61 taxonomic ranks, of which 60 are obtained from TDWG Taxon Rank LSID Ontology[3]. In order to simplify the management of subspecific ranks, an additional class that combines species and taxonomic levels below it was created.

References embody publications in a broad sense including other documented sources of information, for instance minutes of meetings. Bibliographic information can be associated to the reference according to the Dublin Core metadata standard. In biology, author names are often abbreviated when attached to taxon

[2] The TaxMeOn schema is available at http://schema.onki.fi/taxmeon/

[3] http://rs.tdwg.org/ontology/voc/TaxonRank

names. The TaxMeOn model supports the referring system that is typical to biology. Some of the properties used in TaxMeOn are part-specific as the uses of the parts differ from each other. For instance, the property that refers to a vernacular name is only available in the name collection part as it is not relevant in the other parts of the model.

The most distinctive feature of the research part [14] is that a scientific name and taxonomic concepts associated to it are separated, which allows detailed management of them both. In the name collection and species list parts, a name and its taxonomic concepts are treated as a unit. Different statuses can be associated to names, such as validity (accepted/synonym), a stage of a naming process (proposed/accepted) and spelling errors.

The model has a top-level hierarchy that is based on a rough classification, such as the division of organism *classes* and *orders*. Ontologies that are generated using TaxMeOn, can be hung on the top-level classification. A hierarchy is created using the transitive *isPartOfHigherTaxon* relation, e.g. to indicate that the species *forrestii* belongs to the genus *Abies*.

Taxon names that refer to the same taxon can occur as different names in the published species lists and different types of relations (see Table 1) can be set between the taxa. Similarly, research results of phylogenetic studies can be mapped using the same relations. The relations for mapping taxa are divided on the basis of attributes of taxa (*intensional*) or being a member of a group (*ostensive*). If it is known that two taxa have an association which is not specified, a class is provided for expressing incomplete information (see the empty ellipse in Fig. 2). This allows associations of taxa without detailed taxonomic knowledge, and especially between taxa originating from different sources.

Table 1. Mapping relations used in species lists and research results. The three relations can be used as intensional and/or ostensive, using their subproperties.

Relation	**Description**
congruent with taxon	taxonomic concepts of two taxa are equal
is part of taxon	a taxonomic concept of a taxon is included in a taxonomic concept of another taxon
overlaps with taxon	taxonomic concepts of two taxa overlap

In TaxMeOn, a reference (an author name and a publication year) to the original publication can be attached to a name. A complete scientific name is atomised into units that can be combined in applications by traversing the RDF graph by utilising the *isPartOfHigherTaxon* and *publishedIn* relations.

Name collections. Scientific names and their taxonomic concepts are treated as one unit in the name collection, because the scope is in vernacular names. The model supports the usage of multiple languages and dialects of common names. There may be several common names pointing to the same taxon, and typically one of them is recommended or has an official status. Alternative names are expressed defining the status using the class *VernacularNameStatus* and

references related to the changes of a name status can be added. This allows the tracking the temporal order of the statuses. The model for vernacular names is illustrated in Fig. 2.

Species lists. Species lists have a single hierarchy and they seldom include vernacular names. Species lists have more relevance in science than name collections, but they lack information about name changes and a single list does not express the parallel or contradictory views of taxonomy which are crucial for researchers. Synonyms of taxa are typically presented and the taxonomic concept is included in a name like in a name collection. Taxa occurring in different species lists can be mapped to each other or to research results using the relations in Table 1. In addition, a general association without taxonomic details can be used (see Fig. 2).

Biological research results. In biological research results a key element is a taxonomic concept that can have multiple scientific names (and vice versa). Instead of names, taxonomic concepts are used for defining the relations between taxa. The same relations are applied here as in the speies list part (see Table 1). The latest research results often redefine taxon boundaries, for example a split of taxa narrows the original taxonomic concept and the meaning of the name changes although the name itself may remain the same. The new and the old concepts are connected into a temporal chain by instantiation of a change event. In Fig. 3 the concept of the beetle genus Galba is split into the concepts of the Balgus and Pterotarsus. The taxon names are shown inside the ellipses

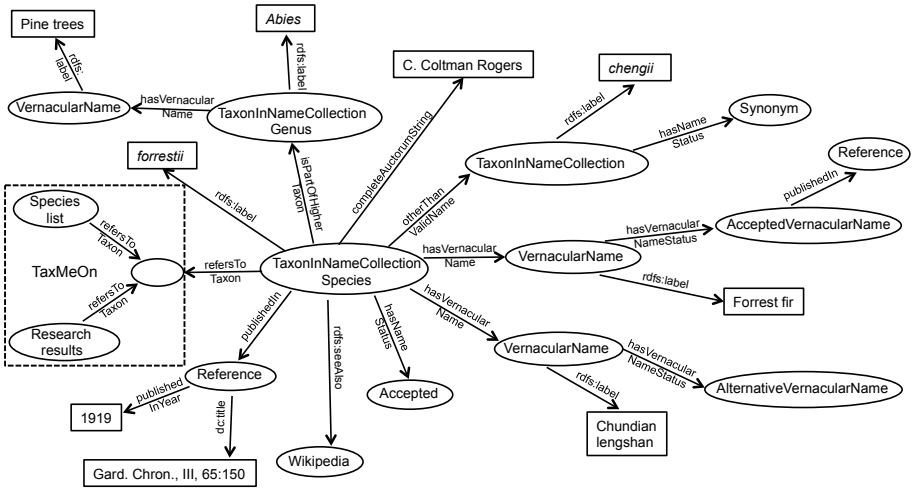


Fig. 2. An example of vernacular names in a name collection. The ellipses represent instances of TaxMeOn classes and literals are indicated as boxes. Other parts of the model are connected to the example taxon in the box with dotted line, in which the empty ellipse illustrates a general representation of a taxon.

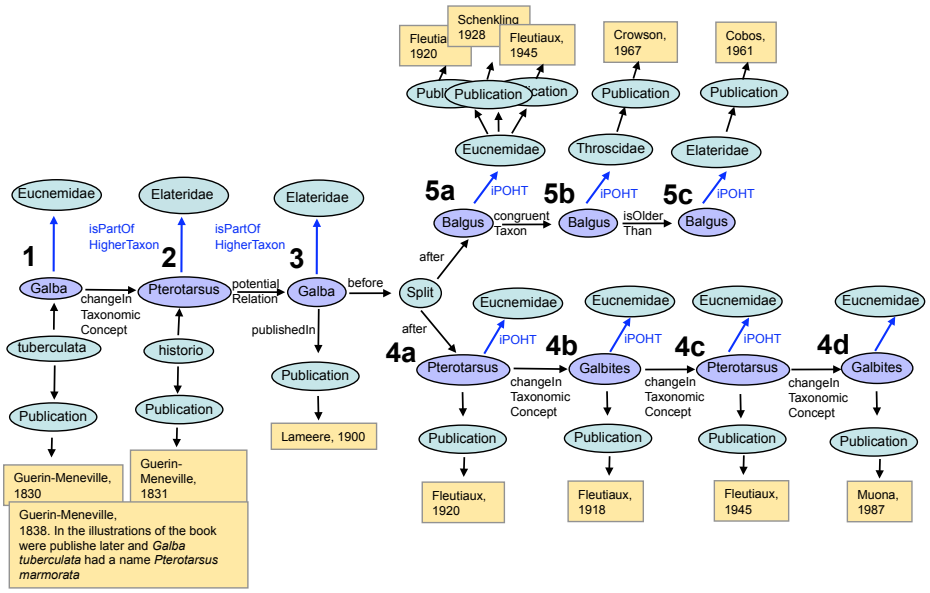


Fig. 3. An example of name changes and taxonomies of eucnemid beetles based on research results. The ellipses represent instances of TaxMeOn classes. Taxonomic hierarchies are expressed with the *isPartOfHigherTaxon* (iPOHT) relations, and the name change series of taxa are illustrated with a darker colour. The following abbreviations are used for the change types: S = Split of taxa, NC = Name change, TCC = Taxon concept change and CH = Change in hierarchy. The meaning of the numbers: 1) The species description of *Galba tuberculata* was originally published in 1830, but the illustrations of the book were published in 1838. However, in the illustrations *G. tuberculata* appeared with the name *Pterotarsus marmorata* (conflicting information). 2) Meanwhile, in 1831, the same taxon was independently described as *Pterotarsus historio* (independent events). 3) Lameere was confused by the two independently published works and changed the name to *Galba* in 1900 (uncertain relation between 2 and 3). 4a) Fleutiaux split the genus *Galba* into two genera. The name *Galba* was changed into *Pterotarsus* as there turned out to be a crustacean genus *Galba* (S, NC,TCC). 4b) Fleutiaux re-examined the genus and concluded that it is new to science and described it as *Galbites* (NC, TCC). 4c) Later Fleutiaux changed his mind and renamed the genus as *Pterotarsus* again (NC, TCC). 4d) Muona discovered that Fleutiaux was originally right and renamed the genus as *Galbites* (NC, TCC). 5a) When *Galba* was split, a part of its species were shifted into the genus *Balgus* that was described as new to science at the same time. *Balgus* was placed in the family Eucnemidae (CH). 5b) And changed into the family Throscidae (CH). This was originally published in a monthly magazine in the 1950's, but the magazines were published as a book in 1967 which is most commonly cited. 5c) *Balgus* was changed into the family Elateridae in 1961 (CH and conflict in publication years).

representing taxonomic concepts in order to simplify the presentation. Other change types are a lump of taxa, a change in taxon boundaries and a change in a hierarchy. These changes lead to the creation of a new instance of a taxonomic concept in order to maintain the traceable taxon history. An instantion of a new concept prevents evolving non-existing name combinations and artificial classifications. For instance, a species name is not associated with a genus name in which it has never been included.

The status of a scientific name may change in time as an accepted name may become a synonym. Multiple statuses can be attached to a name, but according to the nomenclatural rules only one of them is accepted at time. The temporal order of the statuses can be managed according to the same idea as in the name collections part.

4 Use Cases

We have applied the TaxMeOn ontology model to three use cases that are based on different needs. The datasets include a name collection of common names of vascular plants, several species lists of different animal groups and a collection of biological research results of Afro-tropical beetles. The use cases were selected on the basis of the active usage of the data (vernacular names), usefulness to the users (species lists), and the taxonomic challenges with available expertise (scientific names based on research results). The datasets used are depicted in Table 2.

4.1 Collaborative Management of Vascular Plants Names

The biological name collection includes 26,000 Finnish names for vascular plants that are organised into a single hierarchy. A deeply nested hierarchy is not necessary here as the classification used is robust, containing only three taxonomic ranks. The need is to maintain the collection of the common names and to manage the name acceptance process. The number of yearly updates exceeds 1,000. The typical users of the name collection are journalists, translators and other non-biologists who need to find a common name for a scientific name.

The name collection of vascular plants is managed in SAHA[4] [12]. SAHA is a simple, powerful and scalable generic metadata editor for collaborative content creation. Annotation projects can be added into SAHA by creating the metadata schema for the content and loading it into SAHA. The user interface of SAHA adapts to the schema by providing suitable forms for producing the metadata. The values of the properties of the schema can be instances of classes defined in the schema, references to ontologies or literals. The annotations created using SAHA are stored in a database, from which they can be retrieved for use in semantic applications. SAHA also provides a SPARQL endpoint for making queries to the RDF data.

[4] http://demo.seco.tkk.fi/saha/VascularPlants/index.shtml

Table 2. Datasets TaxMeOn has been applied to. Vascular plants are included in the name collection, the false click beetles are biological research results, and all other datasets are based on species lists.

Taxon group	Region	Publ. years	# of taxa
Vascular plants	World	constantly updated	25726
Long-horn beetles (Coleoptera: Cerambycidae)	Scandinavia, Baltic countries	1939, 1960, 1979, 1992, 2004, 2010, 2010	205, 181, 247, 269, 300, 297, 1372
Butterflies and moths (Lepidoptera)	Scandianavia, North-West Russia, Estonia	1962, 1977, 1996, 2002, 2008	313, 256, 265, 4573, 12256, 3244, 3251, 3477
Thrips (Thysanoptera)	Finland	2008	219
Lacewings and scorpionflies (Neuroptera and Mecoptera)	Finland	2008	113
True bugs (Hemiptera)	Finland	2008	2690
Flies (Diptera: Brachycera)	Finland	2008	6373
Parasitic wasps (Hymenoptera: Ichneumoidae)	Finland	1995, 1999, 1999, 2000, 2003	282, 398, 919, 786, 733
Bees and wasps (Hymenoptera: Apoidea)	Finland	2010	1048
Mammals	World	2008	6062
Birds	World	2010	12125
False click beetles (Coleoptera: Eucnemidae)	Afrotropics	–	9 genera

New scientific species names are added by creating a new instance of the *Species* class and then adding the other necessary information, such as their status. Similarly, a higher taxon can be created if it does not already exist, and the former is linked to the latter with the *isPartOfHigherTaxon* relation. SAHA has search facilities for querying the data, and a journalist writing a non-scientific article about a house plant, for example, can use the system for finding a common name for the plant.

4.2 Publishing Species Lists as Ontology Services

The users of species lists are ecologists, environmental authorities and amateurs searching for the correct scientific name occurring in a certain geographical area. In this use case ca. 20 published species lists obtained from the taxonomic database of the Finnish Museum of Natural History[5] containing more than 80,000 names were converted into TaxMeOn ontologies. In addition, seven regional lists of long-horn beetles (cerambycids) with 100 species are available from the years 1936–2010. The various names meaning the same taxon were mapped by an expert. The most common differences between the lists are a shift of a genus for a species, a change in a hierarchy and/or in a name status. Similarly, ca. 150 species of butterfly names from five lists were mapped.

[5] http://taxon.luomus.fi/

Currently, the mapped beetle names are published as services for humans and machines in the ONKI Ontology Service[6] [25]. The ONKI Ontology Service is a general ontology library for publishing ontologies and providing functionalities for accessing them, using ready-to-use web widgets as well as APIs. ONKI supports content indexing, concept disambiguation, searching, and query expansion.

Fig. 4 depicts the user interface of the ONKI server [24]. The user is browsing the species lists of cerambycid beetles, and has made a query for taxon names starting with a string "ab". The selected species *abdominalis* has been described by Stephens in 1831, and it occurs in the species list Catalogue of Palaearctic Coleoptera, published in the year 2010 [15]. The species *abdominalis* belongs to the subgenus and genus *Grammoptera*. The taxonomy of the family Cerambycidae is visualised as a hierarchy tree. The same species also occurs in other species lists, which is indicated by *congruentWithTaxonOst* relation. Browsing the taxa reveals varying taxon names and classifications. For example, the *Grammoptera (Grammoptera) abdominalis* has a subgenus in this example, but the rank subgenus does not exist in the other lists of cerambycid. Also, the synonyms of the selected taxon are shown (*analis*, *femorata*, *nigrescens* and *variegata*).

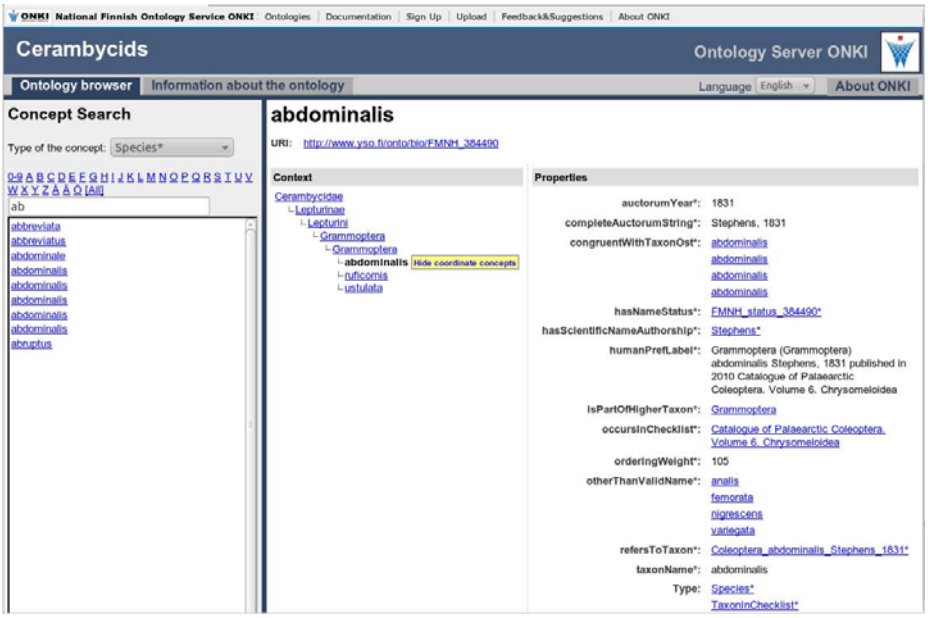


Fig. 4. The species of *abdominalis* shown in the ONKI Browser

The ONKI Ontology Services can be integrated into applications on the user interface level (in HTML) by utilising the ONKI Selector, a lightweight web widget providing functionalities for accessing ontologies. The ONKI API has

[6] http://demo.seco.tkk.fi/onkiskos/cerambycids/

been implemented in three ways: as an AJAX service, as a Web Service, and as a simple HTTP API.

The ONKI Ontology Service contains several ontologies covering different fields and is a part of the FinnONTO project [6] that aims to build a national ontology infrastructure. The Finnish Spatio-temporal Ontology (SAPO) [8], for example, can be used to disambiguate geographical information of observational data. Combining the usage of species ontologies and SAPO, extensive data harmonisation is avoided as both taxon names and geographical names change in time.

4.3 Management of Individual Scientific Names

The use case of scientific names is the Afro-tropical beetle family Eucnemidae, which consists of ca. nine genera that have gone through numerous taxonomic treatments. Also, mistakes and uncertain events are modelled if they are relevant to name changes. For example, the position of the species *Pterotarsus historio* in taxonomic classification has changed 22 times and at least eight taxonomic concepts are associated to the genus *Pterotarsus* [17]. Fig. 3 illustrates the problematic nature of the beetle group in a simplified example. A comparable comparable situation concerns most organism groups on Earth. Due to the numerous changes in scientific names, even researchers find it hard to remember them and this information can only be found in publications of taxonomy. The option of managing individual names is advantageous as it completes the species lists and allows the mapping of detailed taxonomic information to the species lists. For example, environmental authorities and most biologists prefer a simple representation of species lists instead of complicated change series.

5 Discussion

We have explored the applicability of the semantic web technologies for the management needs of biological names. Separating taxonomic concepts from scientific and vernacular names is justified due to the ambiguity of the names referring to taxa. This also enables relating relevant attributes separately to a concept and to a name, although it is not always clear to which of these an attribute should be linked and subjective decisions have to made. The idea of the model is simplicity and practicality in real-world use cases.

The fruitfulness lays in the possibilities to link divergent data serving divergent purposes and in linking detailed information with more general information. For example, a common name of a house plant, a taxonomic concept that appears to be a species complex (a unit formed by several closely related species) and the geographical area can be linked.

The most complex use case is the management of scientific name changes of biological research results. The main goal is to maintain the temporal control of the name changes and classifications. The instantiation of taxon names and concepts lead to a situation in which they are hard to manage when they form a

long chain. Every change increases the number of instances created. Protegé[7] was used for editing the ontologies, although managing names is quite inconvenient because they are shown as an alphabetically ordered flat list, not as a taxonomic hierarchy.

As Protegé is rather complicated for a non-expert user, the metadata editor SAHA was used for maintaining the continuous changes of common names of plants. The simplicity of SAHA makes it a suitable option for ordinary users who want to concentrate on the content. However, we noticed that some useful features are missing from SAHA. The visualisation of a nested hierarchy would help users to compare differing classifications.

In many biological ontologies the 'subclass of' relation is used for expressing the taxon hierarchies. However, in the TaxMeOn model we use the *isPartHigherTaxon* relation instead. If the 'subclass of' relation was used to express the taxonomic hierarchy, a taxon would incorrectly be an instance of the higher taxon ranks, e.g., a species would be an instance of the class *Genus*. This would lead to a situation in which queries for genera also return species.

5.1 Related Work

NCBO BioPortal[8] and OBO Foundry[9] have large collections of life science ontologies mainly concentrating on biomedicine and physiology. The absence of taxonomic ontologies is distinctive which may indicate the complexity of the biological name system. The portals contain only three taxonomic ontologies (Amphibian taxonomy, Fly taxonomy and Teleost taxonomy) and one broader classification (NCBI organismal classification). The taxonomic hierarchy is defined using the *rdfs:subClassOf* relation in the existing ontologies. Taxonconcept.org[10] provides Linked Open Data identifiers for species concepts and links data about them originating from different sources. All names are expressed using literals and the following taxonomic ranks are included: a combination of a species and a genus, a class and an order. Parallel hierarchies are not supported. Geospecies[11] uses the properties *skos:broaderTransitive* and *skos:narrowerTransitive* to express the hierarchy.

Page [19] discusses the importance of persistent identifiers for organism names and presents a solution for managing names and their synonyms on the semantic web. The taxon names from different sources referring to the same taxon are mapped using the *owl:sameAs* relation which is a strong statement. Hierarchy is expressed using two different methods in order to support efficient queries.

Schulz et al. [20] presented the first ontology model of biological taxa and its application to physical individuals. Taxa organised in a hierarchy is thoroughly discussed, but the model is static and based on a single unchangeable taxonomy.

[7] http://protege.stanford.edu/
[8] http://bioportal.bioontology.org/
[9] http://www.obofoundry.org/
[10] http://www.taxonconcept.org/
[11] http://lod.geospecies.org/

Despite recognising the dynamic nature of taxonomy and the name system, the model is not applicable in the management of biological names as such.

Franz and Peet [3] enlighten the problematic nature of the topic by describing how semantics can be applied in relating taxa to each other. They introduce two essentially important terms from philosophy to taxonomy to specify the way, in which differing classifications that include different sets of taxa can be compared. An ostensive relation is specified by being a member of a group and intensional relations are based on properties uniting the group. These two fundamentally different approaches can be used simultaneously, which increases the information content of the relation.

Franz and Thau [4] developed the model of scientific names further by evaluating the limitations of applying ontologies. They concluded that ontologies should focus either on a nomenclatural point of view or on strategies for aligning multiple taxonomies.

Tuominen et al. [23] model the taxonomic hierarchy using the *skos:broader* property, and preferred scientific and common names of the taxa are represented with the property *skos:prefLabel* and alternative names with *skos:altLabel*. The property *rdf:type* is used to indicate the taxonomic rank. This is applicable to relatively simple taxonomies such as species lists, but it does not support expressing more elaborate information (changes in a concept or a name).

The Darwin Core (DwC) [2] is a metadata schema developed for observation data by the TDWG (Biodiversity Information Standards). The goal of the DwC is to standardise the form of presenting biological information in order to enhance the usage of it. However, it lacks the semantic aspect and the terms related to biological names are restricted due to the wide and general scope of the DwC.

The scope of the related work presented above differs from our approach as our focus is on practical name management and retrieval of names.

Research on ontology versioning [10] and ontology evolution [18] has focused on finding mappings between different ontology versions, performing ontology refinements and other changes in the conceptualisation [9,21], and in reasoning with multi-version ontologies [5]. There are similarities in our problem field, but our focus is to support multiple parallel ontologies interpreting the domain differently, not in versioning or evolution of a specific ontology. For example, there is no single taxonomy of all organisms, but different views of how they should be organised into hierarchies.

A similar type of an approach for managing changes and parallel views of concepts has been proposed by Tennis and Sutton [22] in the context of SKOS vocabularies. However, TaxMeOn supports richer ways of expressing information, e.g. for managing changes of taxon names and concepts separately.

5.2 Future Work

The model will be tested using different datasets to ensure its applicability. Currently, the research results part covers animal names, but will be expanded to plant names as well. The lack of user-friendly tools is obvious and the metadata

editor SAHA is planned to be expanded to respond to the needs. Describing evolutionary trees and their information content is a challenging application area as phylogenetics produces name changes.

Acknowledgements. This work is part of the National Semantic Web Ontology project in Finland[12] (FinnONTO, 2003-2012), funded mainly by the National Technology and Innovation Agency (Tekes) and a consortium of 38 organizations. We thank Jyrki Muona, Hans Silfverberg, Leo Junikka and Juhana Nieminen for their collaboration.

References

1. Berendsohn, W.: The concept of "potential taxon" in databases. Taxon 44, 207–212 (1995)
2. Darwin Core Task Group. Darwin core. Tech. rep (2009), http://www.tdwg.org/standards/450/
3. Franz, N., Peet, R.: Towards a language for mapping relationships among taxonomic concepts. Systematics and Biodiversity 7(1), 5–20 (2009)
4. Franz, N., Thau, D.: Biological taxonomy and ontology development: scope and limitations. Biodiversity Informatics 7, 45–66 (2010)
5. Huang, Z., Stuckenschmidt, H.: Reasoning with multi-version ontologies: A temporal logic approach. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 398–412. Springer, Heidelberg (2005)
6. Hyvönen, E., Viljanen, K., Tuominen, J., Seppälä, K.: Building a national semantic web ontology and ontology service infrastructure – the FinnONTO approach. In: Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M. (eds.) ESWC 2008. LNCS, vol. 5021, pp. 95–109. Springer, Heidelberg (2008)
7. ICZN. Opinion 2245 (case 3407) drosophila fallén, 1823 (insecta, diptera): Drosophila funebris fabricius, 1787 is maintained as the type species. Bulletin of Zoological Nomenclature 67(1) (2010)
8. Kauppinen, T., Väätäinen, J., Hyvönen, E.: Creating and using geospatial ontology time series in a semantic cultural heritage portal. In: Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M. (eds.) ESWC 2008. LNCS, vol. 5021, pp. 110–123. Springer, Heidelberg (2008)
9. Klein, M.: Change Management for Distributed Ontologies. Ph.D. thesis, Vrije Universiteit Amsterdam (August 2004)
10. Klein, M., Fensel, D.: Ontology versioning on the Semantic Web. In: Proceedings of the International Semantic Web Working Symposium (SWWS), July 30 – August 1, pp. 75–91. Stanford University, California (2001)
11. Knapp, S., Polaszek, A., Watson, M.: Spreading the word. Nature 446, 261–262 (2007)
12. Kurki, J., Hyvönen, E.: Collaborative metadata editor integrated with ontology services and faceted portals. In: Workshop on Ontology Repositories and Editors for the Semantic Web (ORES 2010), the Extended Semantic Web Conference ESWC 2010, CEUR Workshop Proceedings, Heraklion, Greece (June 2010), http://ceur-ws.org/

[12] http://www.seco.tkk.fi/projects/finnonto/

13. Laurenne, N., Tuominen, J., Koho, M., Hyvönen, E.: Modeling and publishing biological names and classifications on the semantic web. In: TDWG 2010 Annual Conference of the Taxonomic Databases Working Group (September 2010); poster abstract
14. Laurenne, N., Tuominen, J., Koho, M., Hyvönen, E.: Taxon meta-ontology TaxMeOn – towards an ontology model for managing changing scientific names in time. In: TDWG 2010 Annual Conference of the Taxonomic Databases Working Group (September 2010); contributed abstract
15. Löbl, I., Smetana, A.: Catalogue of Palearctic Coleoptera Chrysomeloidea, vol. 6. Apollo Books, Stenstrup (2010)
16. Mayden, R.L.: A hierarchy of species concepts: the denouement in the saga of the species problem. In: Claridge, M.F., Dawah, H.A., Wilson, M.R. (eds.) Species: The Units of Biodiversity Systematics Association Special, vol. 54, pp. 381–424. Chapman and Hall, London (1997)
17. Muona, J.: A revision of the indomalesian tribe galbitini new tribe (coleoptera, eucnemidae). Entomologica Scandinavica. Supplement 39, 1–67 (1991)
18. Noy, N., Klein, M.: Ontology evolution: Not the same as schema evolution. Knowledge and Information Systems 6(4) (2004)
19. Page, R.: Taxonomic names, metadata, and the semantic web. Biodiversity Informatics 3, 1–15 (2006)
20. Schulz, S., Stenzhorn, H., Boeker, M.: The ontology of biological taxa. Bioinformatics 24(13), 313–321 (2008)
21. Stojanovic, L.: Methods and Tools for Ontology Evolution. Ph.D. thesis, University of Karlsruhe, Germany (2004)
22. Tennis, J.T., Sutton, S.A.: Extending the simple knowledge organization system for concept management in vocabulary development applications. Journal of the American Society for Information Science and Technology 59(1), 25–37 (2008)
23. Tuominen, J., Frosterus, M., Laurenne, N., Hyvönen, E.: Publishing biological classifications as SKOS vocabulary services on the semantic web. In: TDWG 2010 Annual Conference of the Taxonomic Databases Working Group (September 2010); demonstration abstract
24. Tuominen, J., Frosterus, M., Viljanen, K., Hyvönen, E.: ONKI SKOS server for publishing and utilizing SKOS vocabularies and ontologies as services. In: Aroyo, L., Traverso, P., Ciravegna, F., Cimiano, P., Heath, T., Hyvönen, E., Mizoguchi, R., Oren, E., Sabou, M., Simperl, E. (eds.) ESWC 2009. LNCS, vol. 5554, pp. 768–780. Springer, Heidelberg (2009)
25. Viljanen, K., Tuominen, J., Hyvönen, E.: Ontology libraries for production use: The finnish ontology library service ONKI. In: Aroyo, L., Traverso, P., Ciravegna, F., Cimiano, P., Heath, T., Hyvönen, E., Mizoguchi, R., Oren, E., Sabou, M., Simperl, E. (eds.) ESWC 2009. LNCS, vol. 5554, pp. 781–795. Springer, Heidelberg (2009)

Publication VII

Nina Laurenne, Jouni Tuominen, Hannu Saarenmaa and Eero Hyvönen. Making species checklists understandable to machines – a shift from relational databases to ontologies. ***Journal of Biomedical Semantics*****, 5, 40, DOI 10.1186/2041-1480-5-40, September 2014.**

Laurenne *et al. Journal of Biomedical Semantics* 2014, **5**:40
http://www.jbiomedsem.com/content/5/1/40

JOURNAL OF BIOMEDICAL SEMANTICS

RESEARCH Open Access

Making species checklists understandable to machines – a shift from relational databases to ontologies

Nina Laurenne[1*†], Jouni Tuominen[1†], Hannu Saarenmaa[2] and Eero Hyvönen[1]

Abstract

Background: The scientific names of plants and animals play a major role in Life Sciences as information is indexed, integrated, and searched using scientific names. The main problem with names is their ambiguous nature, because more than one name may point to the same taxon and multiple taxa may share the same name. In addition, scientific names change over time, which makes them open to various interpretations. Applying machine-understandable semantics to these names enables efficient processing of biological content in information systems. The first step is to use unique persistent identifiers instead of name strings when referring to taxa. The most commonly used identifiers are Life Science Identifiers (LSID), which are traditionally used in relational databases, and more recently HTTP URIs, which are applied on the Semantic Web by Linked Data applications.

Results: We introduce two models for expressing taxonomic information in the form of species checklists. First, we show how species checklists are presented in a relational database system using LSIDs. Then, in order to gain a more detailed representation of taxonomic information, we introduce meta-ontology TaxMeOn to model the same content as Semantic Web ontologies where taxa are identified using HTTP URIs. We also explore how changes in scientific names can be managed over time.

Conclusions: The use of HTTP URIs is preferable for presenting the taxonomic information of species checklists. An HTTP URI identifies a taxon and operates as a web address from which additional information about the taxon can be located, unlike LSID. This enables the integration of biological data from different sources on the web using Linked Data principles and prevents the formation of information silos. The Linked Data approach allows a user to assemble information and evaluate the complexity of taxonomical data based on conflicting views of taxonomic classifications. Using HTTP URIs and Semantic Web technologies also facilitate the representation of the semantics of biological data, and in this way, the creation of more "intelligent" biological applications and services.

Keywords: Scientific name, Taxonomic concept, LSID, HTTP URI, Ontology, Semantic web, Linked data, Species checklist

Background

Research on biodiversity requires integrating data from distributed heterogeneous sources, such as scientific literature, observations, and biomedical resources. Data is often presented using a variety of terms, vocabularies, and languages, which presents a barrier to interoperability and makes data reuse and integration a challenge for both human users and machines.

Scientific names are important for interlinking information about taxa in all fields of the Life Sciences. A taxon is a group of one or more organisms whose members are considered evolutionarily related to one another; a taxon typically has a name and rank, i.e., a species, genus, etc. Taxon names are especially necessary when indexing biological information and cataloguing biodiversity. The nature of names, whether important or problematic, has recently been re-examined by several researchers [1-6].

*Correspondence: nina.laurenne@helsinki.fi
†Equal contributors
[1]Semantic Computing Research Group (SeCo), Department of Media Technology, Aalto University, P.O. Box 15500, 00076 Aalto, Espoo, Finland
Full list of author information is available at the end of the article

Difficulties arise when a particular taxon can be referred to using multiple names, since scientists' opinions differ on how evolutionary units should be organised into classifications. Also, researchers may use the same name with a different meaning when referring to taxa. Well-conducted taxonomic studies may be 250 years old and still useful but in most cases, the perceived boundaries of taxa have been revised several times after the original publication. Contrary to popular belief, a generally agreed-upon, single taxonomy of organisms does not exist, and this fact is directly reflected in the scientific naming system through the various usages of names. For a taxonomist, a scientific name is a label that mirrors an evolutionary hypothesis that is under continuous testing. There will never be a commonly agreed upon single taxonomy and there will always be multiple competing current taxonomic views. Nevertheless, efforts are made to provide usable taxonomies for non-taxonomists.

Checklists are species catalogues where taxa are organised hierarchically according to an author's current view of a classification. The coverage of a species checklist varies from a geographically limited area to a worldwide list, and it typically focuses on a particular organismal group. An author's view of research results is thus inevitably emphasised, which opens the lists to interpretation if they lack sufficient taxonomic details. A regional species list indexes taxa of a given area, but it can also contain additional information. For example, Fauna Europaea [7] and the Atlas of Living Australia [8] provide distribution maps and visualisation tools. The database Encyclopedia of Life (EoL) [9] covers the whole world and has a considerable amount of species information. Also, unlike most resources, it supports multiple classifications since data providers can upload differing taxonomies into the system.

Checklists were previously only published in journals (static lists), but up-to-date checklists (dynamic lists) are increasingly available on the web. For example, the most notable database, Catalogue of Life (CoL) [10], aims to include all known species and currently contains nearly 1,352,112 species from 132 taxonomic datasets (2013 Annual Checklist). The database of zoological names ZooBank [11] currently has 101,777 nomenclatural acts. The Global Biodiversity Information Facility (GBIF) [12] has made an effort to stabilise name usage by setting up a Checklist Bank [13] for storing names and information about them. The widely used Taxonomic Concept Transfer Schema (TCS) [14] specifies the format (XML), in which taxonomic information is presented when exchanging data. Darwin Core (DwC) [15,16], created by Biodiversity Information Standards (TDWG) [17], is a standardised form of presenting biological information. The metadata elements in DwC are not strictly defined as the format and the element values are not fully specified. This means that the interoperability of DwC records is not achieved if the elements are not used in a consistent way. For example, a taxon name may be a literal value or referred to using a URI. Darwin Core Archive (DwC-A) [18] is a data standard for producing a self-contained dataset for sharing species-related data, such as occurrence records and checklists. The CSV (Comma-Separated Values) data files of an archive are organised in a star-like manner, with one core data file and possible extensions, e.g., for vernacular names or distribution data.

The scope of biomedical resources differs from checklists because the focus is on a gene or a cell level. Nevertheless, the name question remains relevant because scientific names are used for linking information. Currently, the National Center for Biotechnology Information (NCBI) [19] provides a single robust consensus hierarchy of taxa constructed by experts, but NCBI ambitiously seeks to build a topology based on monophyletic groups, i.e., taxa derived from a common ancestor. NCBI allows flexibility in the acceptance of informal names and surrogate names can be used when contributing data and searching for taxa [5]. The majority of the submitted DNA sequences do not have a binominal scientific name because specimens are not identified into a species level at the time of submission or only surrogate names are used [5]. The significance of DNA sequence data is increasing due to the rapid development of molecular methods that are applied in constructing evolutionary hypotheses and barcoding biodiversity. Consequently, descriptions of new species based on molecular evidence result in an increased number of species in checklists.

A major source for ambiguity in scientific names is that they change over time. One of the most common types of change concerns a Linnean binominal name combination. The genus of a binominal name changes when a species is moved to another genus. For example, the parasitic wasp species *moscaryi* once belonged to the genus *Tetraconus*, but as a result of a taxonomic revision that synonymised two genera, its new name combination is *Monomachus moscaryi* [20]. Synonymisation happens due to assessments of the identity of types (i.e., typically a physical specimen to which a scientific name is attached). If two or more taxa are lumped, the older name remains valid but with a changed taxonomic circumscription, and the more recent names become its synonyms. Consequently, there is more than one name pointing to the taxon, and the taxonomic concept associated with the older name changes. The opposite situation is the split of taxa, where one taxon is divided into two or more taxa. The divergence between a name and its meaning is characteristic of taxonomy, because a scientific name does not necessarily change despite the fact that taxon boundaries are redefined. Researchers can also classify the same species

in various ways, thus leading to the existence of multiple name combinations.

Berendsohn [21] introduced the concept of a potential taxon, which is a scientific name with information on a circumscription. He proposed the term "secundum" (abbr. "sec") be attached to a scientific name when referring to a particular taxonomic circumscription. This was a concrete suggestion on how to interlink differing taxonomic views while continuing to retain the adequate taxonomic information in databases [22]. Having information on circumscriptions in databases is an improvement, but machine-readable semantics need to be used in order to enhance the machine-processability of taxonomic information.

In this paper we present two models for describing taxonomic information in a machine-processable way. The first model describes species checklists as a relational database and the second one is further developed representation of taxonomic information using Semantic Web technologies. We explore the reasons for moving away from relational databases towards semantic technology, and we also discuss options for managing scientific names as they change over time.

Towards semantic handling of biological names

A biologist understands the semantics of scientific names by reading scientific literature, but computers require explicit identifier systems and data models to process semantics. It is obvious that persistent identifiers for taxa should be used instead of ambiguous name strings to increase the processability of scientific names. Using identifiers allows information to be connected unambiguously, which enables interoperability between systems. Furthermore, there is a need to interlink taxa between the different versions of checklists as they are updated. Otherwise, data indexed using an earlier version of a checklist cannot categorically be found using a later version of the checklist.

Recognising taxa using identifiers

The most commonly used identifiers in biology are Life Science Identifiers (LSID) [23]. An LSID consists of six parts (Figure 1): the first two indicate that the type of URN (Uniform Resource Name) is an LSID, the third part expresses the authority, and the fourth specifies the namespace (which specifies the type of an LSID, e.g., scientific name, living thing, picture, or museum specimen), the fifth points to the object ID, and the optional sixth part is for versioning information. An LSID can be accommodated to a single name or to a set of taxonomic details, depending on its purpose [2,24]. For example, identifiers are given to scientific names in the World Register of Marine Species [25], but in the Catalogue of Life [10] they are given to taxonomic concepts. The Universal Biological Indexer and Organizer (uBio) [26] has 11,106,374 namebank records where LSIDs are used for referring to taxonomic concepts [6]. Also, an RDF (Resource Description Framework) representation [27] is provided but some of the essential information is expressed as literals (a classification, taxonomic rank and a typing of resources) instead of URIs, which hampers machine-processability.

The data carried by an LSID is obtained using a specific resolver. Locating the resolver via the Domain Name System (DNS) of the Internet requires that the resolver be configured in a DNS SRV record (DNS service record) of the domain used as the authority part of an LSID. LSIDs can also be used without a resolver if they are presented as HTTP URIs using an LSID HTTP proxy. According to the TDWG guidelines for using identifiers, an LSID resolver should return metadata about the requested resource in RDF form [27]. The application of LSIDs in the Catalogue of Life is thoroughly discussed by Jones et al. [2]. GBIF has published recommendations for the adoption of LSIDs and HTTP URIs [28,29].

The URN scheme applied to LSIDs is a URI scheme standardised by the Internet Assigned Numbers Authority (IANA) [30]. HTTP is also a URI scheme, but there is a fundamental difference between URNs and HTTP URIs. HTTP URIs are based on the DNS, where the global uniqueness of identifiers is guaranteed by the DNS infrastructure, which also facilitates addressing and retrieving information about HTTP URIs. In contrast to URNs, separate web services are not necessary to manage identifier creation or resolve them for data retrieval because these functions are already available in the infrastructure of the web. As a result, HTTP URIs are used as the identifier mechanism for the Semantic Web and Linked Data [31]. In addition, the form of an HTTP URI is flexible because

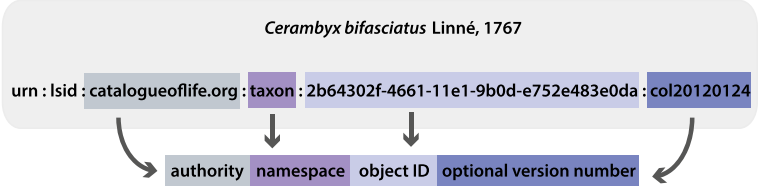


Figure 1 The structure of an LSID. An LSID of a cerambycid beetle species obtained from the Catalogue of Life database.

it does not have strictly defined parts like LSIDs. HTTP URIs allow linking data across the web on the basis of the meaning of concepts that are identified with HTTP URIs, which enables the creation of the Web of Data.

LSIDs were the first attempt to solve the name problem, but due to the rapid development of Semantic Web technologies, the trend now favours standardised web technology. The main differences between LSIDs and HTTP URIs are presented in Table 1. The technology applied does not solve the problem of the divergence between a name and its meaning, but it does provide an appropriate solution for publishing and interlinking data in an interoperable way on the web.

Both LSID- and HTTP URI-based checklists can be published for humans via a user interface and for machines as APIs (Application Programming Interface) to provide access to the data in multiple ways. For example, the user interface can be used to check a valid name for a taxon and browse a classification. The same information can be obtained using a specialised API, but more general query interfaces can also be provided. In Linked Data, an API for reading the RDF description or a human-readable HTML page for a resource is typically provided, as well as a general purpose endpoint service that can be queried using the Semantic Web query language SPARQL. In addition, checklists can be made available as downloadable files [31].

Semantic modelling of taxonomies

On the Semantic Web, taxonomies are represented using RDF resources, i.e., entities with URI identifiers, and explicit relations between them. A relatively new approach is to express taxonomic information as an ontology. The first ontology model for a taxonomic classification was presented by Schulz et al. [34], with taxa organised into a single hierarchy. Franz and Peet [35] and Franz and Thau [36] have offered further insight into the issues of taxonomic ontology modelling. So far, a few taxonomic ontologies have been published in the NCBO BioPortal [37-41] and the ONKI ontology service [42]. The most comprehensive of them is the NCBI Organismal classification [41], which contains more than 352,000 taxa in a single hierarchy. Common to the classifications in the NCBO BioPortal is that the hierarchy is constructed using *subclassOf* (isA) relations and presented in the OBO ontology language [43]. TaxonConcept.org [44] tackles the name problem of taxonomic information in practice and shows how to publish the information as Linked Open Data. It also demonstrates how data from external sources are integrated and investigates how to combine taxonomic concepts with specimen data. However, some of the important information about names are described as literals, e.g., the classification of taxa. Also, the taxonomic change types are not described (split or lump of taxa). The Taxonomic Meta-Ontology TaxMeOn [45,46] aims to respond to the practical needs of managing biological names over time, and it links taxonomic information to names. This meta-ontology differs in that it offers a greater level of detail and supports differing classifications.

An increasing number of ontologies are available and therefore ontology evolution has become an important issue. The world – and our conceptualisation of it – is continually changing, which makes ontology versioning essential [45,47,48]. Existing data that refer to a concept should be kept consistent when its meaning changes or when it is removed from an ontology. Data described using different versions of an ontology then can be integrated by utilising mappings (alignments) between the ontology versions [49]. Khattak et al. [50] document ontology evolution by keeping a log of changes in concepts. Small changes in an ontology are grouped into sets, which can later be used to revert to previous stages. An alternative solution is to recognise concept changes instead of versioning an ontology. Wang et al. [51] show how the changes in concepts and their impacts can be identified automatically by comparing the concepts both extensionally and intensionally in cases where they do not have fixed identifiers.

Table 1 The main differences between LSIDs and HTTP URIs

	Life science identifiers	HTTP URIs
Standardised by	Object Management Group [32]	Internet Engineering Task Force [33]
Reuse existing URI schemes	Defines a new URN subscheme	Uses an existing scheme
Data retrieval/ dereferenceability	Specific resolving service needed	Uses existing web technology (DNS, web servers)
Structure of identifier	Strict	Flexible
Linked Data compatibility	No	Yes

Methods

In order to develop two models for presenting taxonomic information in a machine-processable way, four design principles were applied to satisfy the following conditions:

1. use as few terms as possible to express as much information as possible in the schema of the model. The taxonomic terminology and its usage is established in biology, and the terms are used in consistent way. As few new terms as possible are introduced.
2. focus on a restricted domain, that is, scientific species checklists including all taxonomic

information and excluding any other taxon-related information (e.g., distribution).
3. support information on various levels of granularity, as the source material is heterogeneous in its level of detail.
4. accept all views of taxonomy equally legitimate regardless of the time they were disseminated.

The focus of the models is in representing the taxonomic relations between taxa in a single checklist (classification, synonymy), in different checklists (mapping taxonomic concepts) and in individual versions of a checklist (managing taxonomic changes).

The datasets utilised in the study consist of 20 published species checklists that cover mainly northern European mammals, birds and several groups of insects and assemble ca. 78,000 taxon names (Additional file 1). Two models are applied to the same datasets. Name mappings between the checklists are provided for eight families of papilionoid and hesperioid butterflies.

Results

Taxonomic database

The main elements of the Taxonomic Database (Figure 2) [52] are a binominal scientific name and a taxonomic concept that connects the names that refer to the same taxon. Each concept is identified with a concept LSID. In addition, three other attributes are assigned to the scientific name: 1) a reference to the original publication (author name and year of publication) in which the taxon description was first published, 2) a status of a name indicating its validity in the checklist, and 3) a taxonomic rank expressing level in a hierarchical classification (species, genus, etc.). A taxonomic hierarchy between scientific names is constructed using a hierarchical part-of relation.

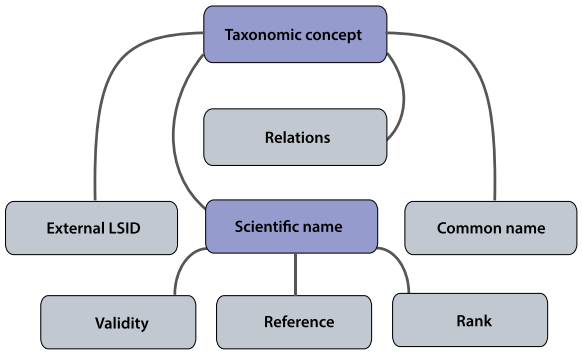


Figure 2 A simplified structure of the relational taxonomic database. The boxes illustrate the tables of the database, and the lines present the relations between them. LSIDs are given to taxonomic concepts and scientific names (illustrated with a darker colour). Taxonomic concepts are linked to each other using the relations described in Table 3 and each concept is linked to a scientific name. External LSIDs and common names are connected to the concepts. An author reference, validity, and a taxonomic rank are assigned to the scientific names.

An LSID that is obtained from an external source can be assigned to a taxon concept as an attribute. Common names in multiple languages can be connected to the concept, but no taxonomic rank can be specified for them. In order to recognise the orthographic variants of scientific names, LSIDs are accommodated to the names as well.

A new LSID is given to a concept if it changes, such as a taxonomic change, an addition or removal of a synonym, or a change in relations between taxa. An LSID is assigned to a new taxon when added to a dynamic checklist. LSIDs are versioned in the case of minor changes using the optional part of the identifier. The decision whether to create a new object identifier of an LSID or a new version is made by a maintainer.

Taxa can be searched using a complete or partial scientific name via a user interface, and the system returns a currently valid name and its synonyms. If the taxon is found in other checklists, their interrelations are also described. The information is also provided as an RDF representation for machine consumption. Only the latest versions of dynamic checklists can be seen in the system. However, older ones are stored internally in the database.

Taxonomic concepts are linked on the basis of their equivalence at a species level, but at higher levels the alignment of taxa is based on the species content. For instance, two species that have the same name and the same authorship citation are linked as congruent by default, but two genera are linked as congruent only if the species belonging to the genera are the same. The reasons for treating species and taxa above the species level differently are debated in the Discussion.

Taxonomic meta-ontology

TaxMeOn is an ontology schema for biological names, and here we present the part that describes species checklists. The model is based on RDFS (RDF Schema) and some features of OWL (Web Ontology Language); it contains 12 classes with 49 subclasses (excluding 61 subclasses of the class *TaxonomicRank*) and 28 properties. The core classes and their relations are illustrated in Figure 3.

The class *TaxonInChecklist* represents both a scientific name and its concept. The relation *rdfs:label* expresses the unominal name of a taxon which is 1) the last epithet of a name combination, or 2) a name of a taxon at higher levels, e.g., a family. The taxonomic hierarchy is constructed using the relation *isPartOfHigherTaxon*.

The author references are presented in two ways:

1. The property *hasScientificNameAuthorship* expresses the author of the original publication (if the

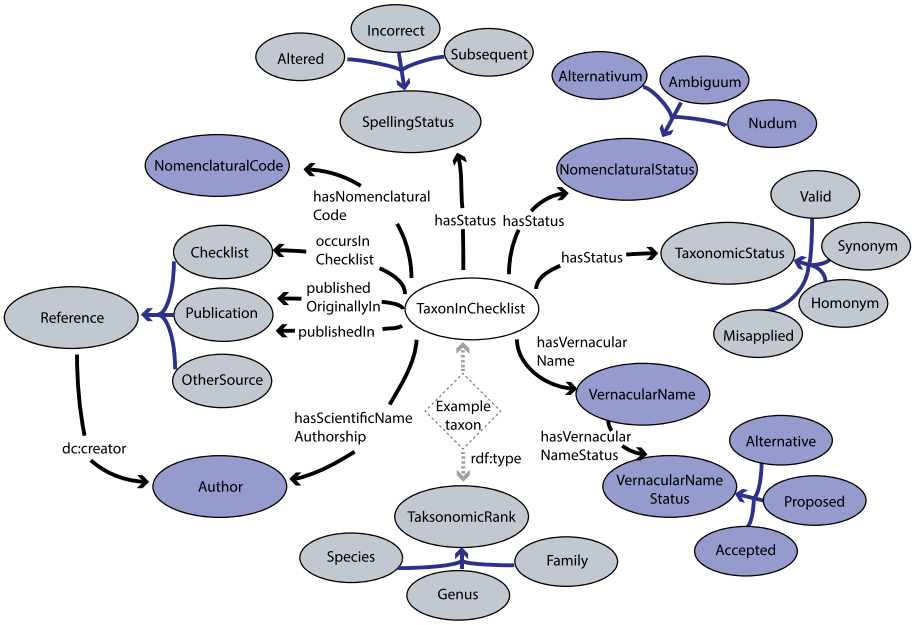


Figure 3 The core classes of the taxonomic meta-ontology. The classes are illustrated with ellipses (colours are to improve the readability of the figure). The arrows indicate relations (properties) between the classes. The subclass relations are indicated with lighter-coloured arrows and a few examples of the subclasses. To demonstrate how the TaxMeOn model is applied, an example taxon depicted using dotted lines is illustrated. The example taxon is an instance of the class *TaxonInChecklist* and of a specific taxonomic rank. The properties associated with the example taxon are marked with dotted-line arrows. The properties with literal values are not shown in the figure.

full reference of the original publication is not provided in a checklist).

2. The properties *publishedIn* and *publishedOriginallyIn* refer to the publication.

The way the taxonomic authority information is worded differs between zoology and botany. Author names are often abbreviated in diverse ways in zoology; for example, both L. and Linn. stand for Linnaeus. In botany, the abbreviations are standardised, but if a species is shifted into another genus, a new author name is catenated into the author reference (unlike in zoology). For instance, Linnaeus first described the species *Bassia scoparia* in the genus *Chenopodium* and later A.J. Scot shifted it into the genus *Bassia*. The order of multiple authors comes out in the literal, i.e., (L.) A.J. Scot.

A binominal name combination of a species with a reference to the original author (e.g., *Arhopalus ferus* (Mulsant, 1839)) is formed by traversing the RDF graph where a genus name is obtained using the *isPartOfHigherTaxon* relation and the other parts of the name from the literals. The literal *completeTaxonName* is for facilitating the usage of the model for humans, and is generated from a genus name, a species name, and an author reference. Dublin Core attributes [53] are supported (e.g., bibliographical details). Figure 4 presents an example of the species *Arhopalus ferus* which was described by Mulsant in 1839 and is a valid name. The same RDF example as Turtle [54] presentation is in Additional file 2.

In Figure 3, the relation *hasStatus* is associated with the class *TaxonInChecklist* and indicates: 1) the nomenclatural status of a name (*nomen alternativum*, *nomen correctum*, etc.), 2) the orthographic variants (altered spelling, incorrect spelling, etc.), and 3) the current opinion of a taxonomic concept (valid, synonym, etc.). Modelling the changes is further discussed in the Discussion. Other important properties and their explanations are listed in Table 2.

The taxonomic concepts are mapped using the relations described in Table 3. An additional relation *isAssociatedWithTaxon* is provided for linking concepts in taxonomically unresolved cases. The relation describes an undetermined connection between taxa, which is useful if deeper expertise is not available when mapping the concepts.

The taxa can be mapped to an external source as shown below, where the genera *Arhopalus* are mapped congruently between two checklists.

```
@prefix cerambycids: <http://www.yso.fi/onto/cerambycids/>.
@prefix taxmeon: <http://www.yso.fi/onto/taxmeon/>.
cerambycids:p2090 taxmeon:congruentWithTaxonInt
   <urn:lsid:catalogueoflife.org:
   d782a602-29c1-102b-9a4a-00304854f820:col2012acv16>.
```

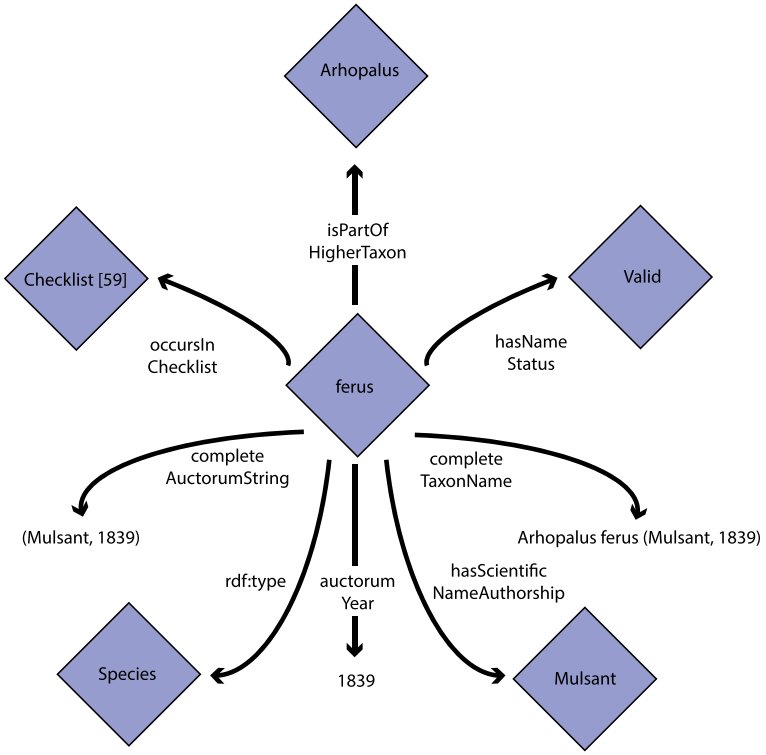


Figure 4 Core taxonomic information represented according to the taxonomic meta-ontology. *Ferus* is described by Mulsant in 1839 and it belongs to the genus *Arhopalus*.

Table 2 The core properties of the Taxonomic Meta-Ontology and their explanations

Property	Explanation
Citation-related properties	
occursInChecklist	Reference to a species checklist
auctorumYear	The year of original publication
completeAuctorumString	Author name(s) expressed according to the established practices of taxonomy
Name-related properties	
hasNonvalidName	Expresses synonyms, homonyms and orthographic variants of a valid scientific name
hasVernacularName	The common name equivalents for the scientific names
hasNomenclaturalCode	Specifies the set of rules that are applied (ICN [55] or ICZN [56])
hasVernacularNameStatus	Expresses whether a common name is accepted or an alternative one
rdf:type	Expresses the hierarchical level in a classification. The ranks are obtained from TDWG Taxon Rank LSID Ontology [57]. Every taxon is an instance of a specific taxonomic rank and the class *TaxonInChecklist* (Figure 3).

See other properties in the Results section, in subsection Taxonomic Meta-Ontology.

The URI of the scientific name and its concept (*TaxonInChecklist*) is duplicated when there is a taxonomic, nomenclatural, or hierarchical change. In this way, a particular taxon can be explicitly referred to at a particular time. The old and the new URIs are connected with the relations described in Table 3. Temporal management is based on the time stamps of scientific names' taxonomic status in dynamic checklists. In static checklists, the temporal order of the taxon instances is traced by the publication year of the checklist.

Two examples of concept mapping and taxonomic changes are presented below. Each scientific name is given a new URI in each static checklist. Different URIs for the same scientific name enable the presentation of alternative classifications and different sets of taxonomic details. The first example presents four cases presented in static checklists:

1. Two species of long-horn beetles, *pubescens* Fabricius, 1787 and *revestita* Linnaeus, 1767 belong to the genus *Leptura* Linnaeus, 1758 in the checklist that was published in 1992 [58].
2. Both species belong to the genus *Pedostrangalia* Sokolow, 1758 in the checklist published in 2011 [59].

Table 3 The relations used for mapping underlying taxonomic concepts

Relation between taxa	Intensive	Ostensive	Notation	Properties
Congruent	Share the same characters	Share the same species	$A = B$	Symmetric, transitive
Part of	All characters of a taxon are included in another taxon	All species are included in another taxon	$A \subset B$	Non-symmetric, transitive
Overlap	At least one character is shared between taxa, but not all of them	At least one species is shared between taxa, but not all of them	$A \cap B \neq \emptyset, A \neq B$	Symmetric, non-transitive

The division into intensional and ostensive relations [35] is only available in TaxMeOn (not in the Taxonomic Database).

3. The species *L. aethiops* Poda, 1761 remained in the genus *Leptura* while two other species were shifted in 2011 [59].
4. *Pedostrangalia* was a synonym for *Leptura* in 1992 [58].

The corresponding RDF representation is presented in Additional file 3.

The second example describes a fictitious dynamic checklist with three artificial taxa. The species *bus* and *cus* belonged to the genus *Aus* in 2012. Later, these two species were synonymised and *bus* remained a valid name while *cus* became its synonym. The URIs of the scientific names are duplicated in order to: 1) preserve the name combinations of the genus *Aus* (i.e., the lower-level classifications), and 2) present a change in taxonomic concepts and in status of the species *bus* and *cus*. The corresponding RDF representation is presented in Additional file 4.

The checklists are managed using the scalable generic metadata editor SAHA [60], but more complex taxonomic information of the scientific names is managed using the ontology editor Protégé [61]. The species ontologies are accessible with several user interfaces and APIs via the Finnish Ontology Library Service ONKI [42,62]. The ONKI browser is used for searching and browsing taxa, finding currently valid names, and tracing the temporal changes in scientific names. The ONKI service also provides an autocompletion widget which can be integrated into user applications, e.g., a content management system. ONKI provides HTTP and SOAP APIs for programmatic access and a SPARQL endpoint for querying the ontologies. The checklists in ONKI are the same as in the Taxonomic Database described earlier.

The HTTP URIs were generated for the data resources in the following form: http://www.yso.fi/onto/CHECKLIST_ID/LOCAL_ID where CHECKLIST_ID is a human-readable identifier for a checklist (or a group of checklists, if there is more than one checklist about the same group) and LOCAL_ID is a local identifier for a resource (e.g., scientific name, taxonomic status). Similarly, the URIs of the authors have namespace, with the CHECKLIST_ID replaced with the string "author". The URIs of TaxMeOn are constructed in the same way, but the CHECKLIST_ID is replaced with the string "taxmeon". LOCAL_ID is in the form "p[NUMBER]", where NUMBER is a randomly generated unique identifier for the checklist data. For the authors and TaxMeOn, the LOCAL_ID is human-readable. The number of RDF triples after the data conversion (TaxMeOn) is over 1,2 million. The details are presented in Additional file 1.

TaxMeOn is applied in a broader context as one of the use cases of the European research program, the "Environmental Observation Web and its Service Applications within the Future Internet (ENVIROFI)" [63] which aims to harmonise biodiversity observation data gathered from heterogeneous sources.

Discussion

Identifiers should not embed semantics according to the recommendations of GBIF [28,29], a practical approach to ensure the persistence of the identifiers should the concepts change. In practise, it is helpful if URIs are intuitively understandable to some degree when reading RDF. Here, human-readable checklist identifiers are embedded in the namespace of the URIs in the data, which is justified because the namespaces are permanent. The local names of the URIs, however, do not carry meaning. The identifiers of the classes and properties in ontology models and schemas are typically human-readable, as is the case in TaxMeOn.

The HTTP URIs used in the data and in TaxMeOn act as locators for relevant metadata, that follows the best practices of Linked Data [31]. The metadata is presented as an HTML page to humans and in RDF format to machines via content negotiation.

Comparison of the two models

The differences between the Taxonomic Database and the Taxonomic Meta-Ontology are summarised in Table 4. The Taxonomic Database is a relational database, and therefore its structure is strictly specified in a database schema. The advantage of RDF-based TaxMeOn is that it can easily be extended by adding new classes and properties. Global identifiers (URIs) are given to taxa in TaxMeOn which allows publishing them as Linked Data and linking and re-using heterogeneous data on the web. TaxMeOn can also be utilised via standard SPARQL query language and additional APIs. In contrast to the RDF model, linking other datasets to the Taxonomic Database

Table 4 A comparison of the features of the taxonomic database and the taxonomic meta-ontology

	Taxonomic database	TaxMeOn
Technology		
Structure easily extensible	No	Yes
Global linkability to other contents	No	Yes
Public interfaces	Simple search API, LSID resolver	HTTP and SOAP APIs, Linked Data, SPARQL endpoint
Need of a resolver	Yes	No
Content editing	Web interface	SAHA [60], Protégé [61]
Content		
Granularity of taxonomic information	Low	High
Linking additional scientific publications	No	Yes
Treatment of botanical and zoological names	Identical	Not identical
Semantics applied to author names	No	Yes
Tracking temporal changes	Publication year of a checklist	Versioning of checklists (static) and duplication of taxa (dynamic)

or re-using its data is not straightforward because the data can only be accessed with a separate LSID resolver and a simple search API. The datasets of TaxMeOn can be edited with standard RDF tools, such as ontology editors, whereas the Taxonomic Database is managed with its own web interface. TaxMeOn supports more detailed taxonomic information than the Taxonomic Database, for example nomenclatural treatments. It also allows linking taxa to additional scientific publications and applying semantics to authors instead of presenting them as simple strings. Moreover, TaxMeOn provides versatile methods for managing dynamic checklists by representing temporal changes of taxonomic concepts.

Managing changes in time

In the Taxonomic Database, the goal was to create connections between the scientific names of published checklists where the timeline is evident due to the year of publication. Less emphasis was placed on dynamic lists. However, evincing temporality is achieved by tracking changes in dynamic lists, an activity that requires: 1) keeping a log of taxa removals and additions, 2) creating a new version of a checklist when taxa are removed or added, and 3) linking older LSIDs to the new ones. An original link to a genus should be kept if a species is shifted into another genus.

Also in TaxMeOn, the versioning of a static checklist is the solution for managing names over time given its simplicity in comparison to modelling the changes (Figure 5). Consequently, a large number of URIs are created, which is impracticable for a maintainer if special tools are not developed. Updating taxonomic changes in a dynamic checklist requires the duplication of the URIs at the species and genus level so that the situation before and after can be presented and interlinked. This step is especially necessary if there is a change in a taxonomic concept. The whole upper classification is not duplicated because that would generate a large number of URIs, and here we are more interested in names than classifications.

A machine does not understand that there was a taxonomic change if the change is not modelled. Two alternative ways of describing the changes are demonstrated in Figure 5. One approach is to present a change in a taxonomy, classification, or nomenclature by forming a class that describes the change type (Figure 5A). The situation is described before and after the change, and the two instances are connected with relevant relations (Table 3). In this way, it is possible to refer to a taxonomic concept at a particular time. An alternative approach is to represent the relations as instances (Figure 5B). The relations are ordered temporally by assigning them a time stamp. If the URIs assigned to the concepts are not duplicated, then it is not possible to refer to a taxonomic concept at a particular time. This might be practical in some cases, because a new URI is assigned only to genuinely new information (new hierarchical relations). The former alternative is included in TaxMeOn, and the latter can be used if the model is extended with an additional class that describes the relations between taxa.

Mapping taxonomic concepts

A species checklist is an understandable way of presenting information to non-taxonomists, but unfortunately only a small proportion of species are catalogued, and they cover only limited geographical areas. Moreover, the information is often insufficient because name combinations are not necessarily listed. Cross-linking taxon names between checklists helps a user to piece together the changes in scientific names and determine the approximate number of taxonomic treatments (none vs. many). Linking higher taxa between checklists is rather artificial because the taxonomic concepts are seldom referenced. The problem is therefore how to reconcile the differing classifications of regional checklists. A pragmatic option is to compare the species included in a higher taxon. However, this approach fails to distinguish taxonomy and regionality, leading to a situation where the occurrence of a new species in a certain area changes the existing relations between the higher taxa of checklists.

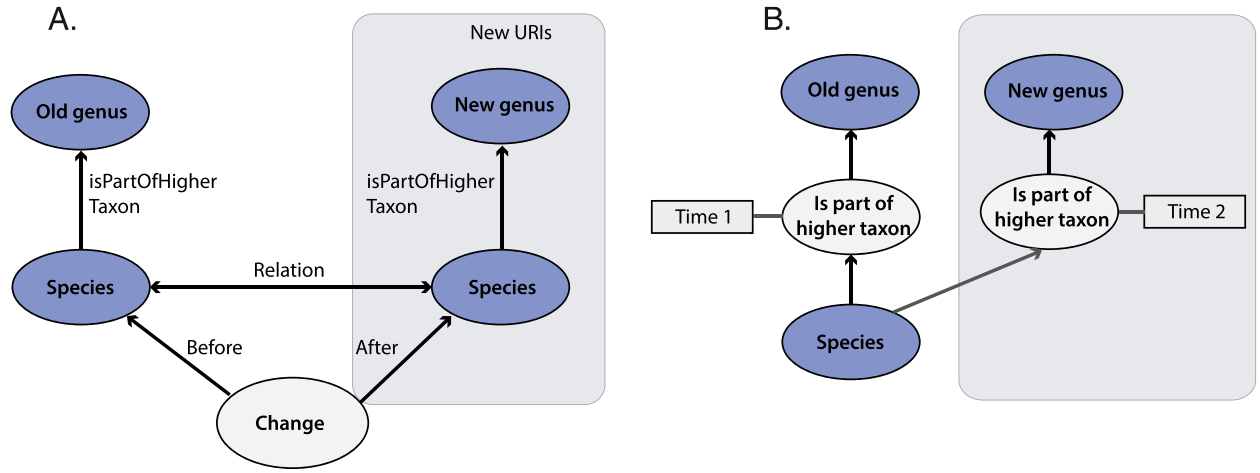


Figure 5 Taxonomic changes in relation to time presented in RDF. (A) The change is modelled as an instance. **(B)** The relation is modelled as an instance. The instances are depicted as lighter-coloured ellipses and literals as rectangles.

The challenges of concept mapping have been discussed by many researchers [2,35,36,64], and it is suggested that it should be stated whether comparisons are based on being a member of a group or on characters that unite the group [35]. In the Taxonomic Database, higher taxa are not only aligned on the basis of underlying taxonomic concepts, but the occurrence of species are also taken into account due to the lack of information about taxonomic concepts in checklists. Higher taxa of the Taxonomic Database are not mapped with the CoL's taxa identifiers because only the part-of relation could be used (because a regional species list is always part of a worldwide list). Instead, external identifiers (CoL) are treated as additional information about the taxonomic concepts. Despite the discrepancy between taxonomy and regionality, a non-taxonomist is more likely to be interested in the species inhabiting a certain geographical area than in those found in the entire world. On the other hand, a maintainer decides how the model is applied. The taxa in both models are mapped equivalently, but TaxMeOn supports more than one way of expressing a relation between taxa (Table 3), which benefits users with differing needs and levels of expertise.

Franz and Peet [35] present how phylogenetic relationships are described using ostensive (i.e., based on being a member of a group) and intensional (i.e., based on characters) relations simultaneously, which increases the semantic precision of the relations between the concepts. In species checklists, there is no satisfactory solution to defining relations at the species level. If ostensive relations are used, there is an assumption that the species have subspecies; however, most species do not have any subspecies. Applying intensional relations assumes that the circumscriptions are known; species lists lack the information on circumscriptions. We decided to use ostensive relations as our default when mapping the concepts at the species level because the nature of the checklist can be interpreted as ostensive because they present a classification. However, intensional relations can be set if there is information about the underlying taxonomic concepts. The comparison of higher taxa (above the species level) is always based on the species (see the discussion of the Taxonomic Database above). The use of ostensive relations (Table 3) differs slightly from Franz and Peet's [35], which is explained by the difference of the data (phylogenies vs. species checklists).

Linking the taxonomic concepts automatically is a quick way of handling datasets. Automatic mapping immediately links new content to existing without time-consuming work by experts that could be done later. A general taxon class (*TaxonGeneral*) represents a taxon at a high level of abstraction, and an instance of it is generated for all taxa. If the taxa share the same name and authorship, then they will be automatically mapped to the same instance of the class *TaxonGeneral*. The idea is to keep the machine-generated mappings separate from the manual ones. The advantage is that if the mappings are used in information retrieval, then search results can be classified according to reliability. Mistakes generated in automated work are inevitable, but most links are likely to be correct due to the non-specific nature of the class *TaxonGeneral*. Different levels of abstraction increase a model's flexibility. For instance, the International Federation of Library Associations and Institutions' (IFLA) [65] Functional Requirements for Bibliographic Records (FRBR) entity-relationship model [66], which is used in online library catalogues, represents the products of intellectual or artistic endeavour at four levels of abstraction.

Challenges

Detailed information is considered more reliable and therefore more likely to be linked to other content than vague information. However, most taxonomic information in checklists is inaccurate in one way or another. Therefore the data model should support the expression of information at various levels of detail, resulting in the complexity of an ontology model. For instance, a taxonomic author citation can include a set of bibliographical details or it simply can be an abbreviation of a name. Our aim was to create a practical model that suits diverse situations, but there is a clear trade-off between practicality and complexity. Combining the scientific name and its concept into a single unit in TaxMeOn increases simplicity but decreases the granularity of information.

The biggest obstacle in using Semantic Web technologies is the lack of suitable tools. Few ontology editors are available. The most commonly used editor is Protégé, which is not practicable in this case because taxonomic classifications cannot be viewed hierarchically unless the *rdfs:subClassOf* relation is used.

In the real world, scientists who study the evolutionary relationships of organisms are often unaware of the advance of biodiversity informatics, or they simply ignore it because they evaluate the usefulness of available resources on the basis of content. Misleading or insufficient information in databases that is copied from one place to another does not encourage scientists to contribute or follow best practices. Taxonomists cannot be expected to follow what happens in biodiversity informatics because it might not be their field of interest. However, it would be very helpful if they were willing to report mistakes in content, though it would be frustrating if the corrections were not made. One can debate whether harmonising names is realistic due to the fact that scientific names are constantly changing. However, applying semantics to the content better enables the presentation of parallel views reflecting the nature of research. Databases and ontologies might not be useful for taxonomists because they rely on scientific publications, and are familiar with their own subject. Regardless, their input is fundamentally important for non-taxonomists, because the need exists for reliable taxonomic information. In general, maintaining and updating ontologies is complex compared to databases. The work is worthwhile, though, because it facilitates interoperability and the semantically enriched processing of content, and brings expert knowledge into wider use in the environmental and biological sciences.

Conclusions

Semantic Web technologies provide a suitable way to describe species checklists because they enable the compatibility with Linked Data. This compatibility is advantageous when reusing and integrating data as well as deepening the level of biological information. Linked Data efficiently prevents the formation of silos, where distributed information is not interlinkable. The advantages of using a Semantic Web approach are presented in Table 4.

Linked Data increases the utility of data gathered from multiple sources, as their reliability is easier to evaluate. For example, the existence of multiple classifications usually indicates that a taxonomic group is complex and many opinions of it exist. Traditional databases are not compatible as such with Linked Data, and they tend to be used internally by organisations rather than shared. The structure of a relational database has to be strictly specified in advance because it cannot be easily changed later, unlike Linked Data, which is more extensible.

The next challenge is to develop a model that addresses both zoological and botanical nomenclatures that are independent of one another and separated by distinct features. We aim to develop an ontology model that covers both nomenclatures without losing the practicality. Applying Semantic Web technologies is a promising step in enhancing the linkability of biological contents and distributing environmentally important information.

Availability of supporting data

The datasets are accessible in the Taxonomic Database, http://taxon.luomus.fi, and in the ONKI Ontology Service, http://onki.fi. The ontology schema of the TaxMeOn model is available at: http://schema.onki.fi/taxmeon/.

Additional files

Additional file 1: Datasets included in the study.

Additional file 2: Core taxonomic information of a checklist expressed in RDF.

Additional file 3: Alternative classifications in static checklists expressed in RDF.

Additional file 4: A synonymisation of taxa in a dynamic checklist expressed in RDF.

Abbreviations

LSID: Life science identifier; HTTP URI: Hypertext transfer protocol uniform resource identifier; TaxMeOn: Taxonomic meta-ontology; EoL: Encyclopedia of life; CoL: Catalogue of life; GBIF: Global biodiversity information facilities; TCS: Taxonomic concept transfer schema; XML: Extensible markup language; DwC: Darwin core; TDWG: Biodiversity information standards; DwC-A: Darwin core archive; CSV: Comma-separated values; NCBI: National center for biotechnology information; URN: Uniform resource name; uBio: Universal biological indexer and organisator; RDF: Resource description framework; DNS: Domain name system; DNS SRV record: DNS service record; IANA: Assigned numbers authority; API: Application programming interface; HTML: Hypertext markup language; SPARQL: SPARQL protocol and RDF query language; NCBO BioPortal: National center for biomedical ontology BioPortal; OBO ontology language: Open biomedical ontologies ontology language; RDFS: RDF schema; OWL: Web ontology language; ICN: International code of nomenclature for algae, fungi, and plants; ICZN: International code of

zoological nomenclature; SOAP: Simple object access protocol; ENVIROFI: The environmental observation web and its service applications within the future internet; IFLA: International federation of library associations and institutions; FRBR: Functional requirements for bibliographic records.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
NL was responsible for the content of the Taxonomic Database and designed the Taxonomic Meta-Ontology with JT. HS designed the structure of the Taxonomic Database of the Finnish Museum of Natural History. EH conceived the study and participated in its design and coordination and helped draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements
The development of the Taxonomic Database was funded by the Nordic Council. The Taxonomic Meta-Ontology has been developed as part of the FinnONTO project, which built a national Semantic Web infrastructure in Finland. Nationally important ontologies are made available as Linked Data in the FinnONTO project. We thank Arto Mertaniemi, Jyrki Muona and Hans Silfverberg for collaboration. The two anonymous reviewers are acknowledged for their comments.

Author details
[1]Semantic Computing Research Group (SeCo), Department of Media Technology, Aalto University, P.O. Box 15500, 00076 Aalto, Espoo, Finland. [2]Digitarium, University of Eastern Finland, P.O. Box 111, 80101 Joensuu, Finland.

Received: 14 April 2013 Accepted: 26 August 2014
Published: 8 September 2014

References

1. Patterson DJ, Cooper J, Kirk PM, Pyle RL, Remsen DP: **Names are key to the big new biology.** *Trends Ecol Evol* 2010, **25**(12):686–691.
2. Jones AC, White RJ, Orme ER: **Identifying and relating biological concepts in the catalogue of life.** *J Biomed Semantics* 2011, **2:**7.
3. Parr CS, Guralnick R, Cellinese N, Page RDM: **Evolutionary informatics unifying knowledge about the diversity of life.** *Trends Ecol Evol* 2012, **27**(2):94–103.
4. Segers H, de Smet WH, Fischer C, Fontaneto D, Michaloudi E, Wallace RL, Jersabek CD: **Towards a list of available names in Zoology, partim Phylum Rotifera.** *Zootaxa* 2012, **3179:**61–68.
5. Federhen S: **The NCBI taxonomy database.** *Nucleic Acids Res* 2012, **40**(Database issue):D136–D143.
6. Sarkar IN: **Biodiversity informatics: organizing and linking information across the spectrum of life.** *Brief Bioinform* 2007, **8**(5):347–357.
7. **Fauna Europaea.** [http://www.faunaeur.org]
8. **Atlas of living Australia.** [http://www.ala.org.au]
9. **Encyclopedia of life.** [http://eol.org]
10. **Catalogue of life.** [http://www.catalogueoflife.org]
11. **ZooBank.** [http://iczn.org/content/about-zoobank]
12. **Global Biodiversity information facility (GBIF).** [http://www.gbif.org]
13. **Checklist bank.** [https://github.com/gbif/checklistbank]
14. Taxonomic Names and Concepts Interest Group: **Taxonomic concept transfer schema.** 2005. [http://www.tdwg.org/standards/117]
15. Darwin Core Task Group: **Darwin core.** 2009. [http://rs.tdwg.org/dwc]
16. Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D: **Darwin core: an evolving community-developed biodiversity data standard.** *PLoS ONE* 2012, **7:**e29715.
17. **Biodiversity information standards (TDWG).** [http://www.tdwg.org]
18. Remsen D, Döring M, Robertson T: **GBIF GNA profile reference guide for darwin core archive, core terms and extensions.** Tech. rep., Global Biodiversity Information Facility (GBIF), Copenhagen, Denmark, 2011. [Version 1.2, released on 1 April 2011]
19. **NCBI national center for biotechnology information.** [http://www.ncbi.nlm.nih.gov]
20. Johnson NF, Musetti L: **Genera of the parasitoid wasp family monomachidae (hymenoptera: diaprioidea).** *Zootaxa* 2012, **3188:**31–41.
21. Berendsohn WG: **The concept of "potential taxa" in databases.** *Taxon* 1995, **44:**207–212.
22. Berendsohn WG: **A taxonomic information model for botanical databases the IOPI Model.** *Taxon* 1997, **46:**283–309.
23. Object Management Group (OMG): **Life sciences identifiers final adopted specification.** 2004. [http://www.omg.org/cgi-bin/doc?dtc/04-05-01]
24. Page RDM: **Taxonomic names, metadata, and the semantic web.** *Biodiversity Inform* 2006, **3:**1–15.
25. **World register of marine species.** [http://www.marinespecies.org]
26. **uBio.** [http://www.ubio.org/]
27. TDWG Globally Unique Identifiers Task Group (GUID): **TDWG life science identifiers (LSID) applicability statement.** 2007. [http://www.tdwg.org/fileadmin/subgroups/guid/LSID_Applicability_Statement_draft.pdf]
28. Cryer P, Hyam R, Miller C, Nicolson N, Tuama EO, Page R, Rees J, Riccardi G, Richards K, White R: **Adoption of persistent identifiers for biodiversity informatics: Recommendations of the GBIF LSID GUID task group, 6. November 2009.** Tech. rep., Global Biodiversity Information Facility (GBIF), Copenhagen, Denmark, 2010. [Version 1.1, last updated 21 Jan 2010]
29. Richards K, White R, Nicolson N, Pyle R: **A beginner's guide to persistent identifiers.** Tech. rep., Global Biodiversity Information Facility (GBIF), Copenhagen, Denmark 2011. [Version 1.0, released on 9 February 2011]
30. Internet Assigned Numbers Authority (IANA): **Uniform resource identifier (URI) schemes.** [http://www.iana.org/assignments/uri-schemes/uri-schemes.xhtml]
31. Heath T, Bizer C: *Linked Data: Evolving the Web into a Global Data Space.* Palo Alto, California: Morgan & Claypool; 2011. [Synthesis Lectures on the Semantic Web: Theory and Technology]
32. **Object management group (OMG).** [http://www.omg.org]
33. **Internet engineering task force (IETF).** [http://www.ietf.org]
34. Schulz S, Stenzhorn H, Boeker M: **The ontology of biological taxa.** *Bioinformatics* 2008, **24**(13):i313–i321.
35. Franz NM, Peet RK: **Towards a language for mapping relationships among taxonomic concepts.** *Syst Biodivers* 2009, **7:**5–20.
36. Franz NM, Thau D: **Biological taxonomy and ontology development: scope and limitations.** *Biodivers Inform* 2010, **7:**45–66.
37. **NCBO BioPortal.** [http://bioportal.bioontology.org]
38. **Amphibian taxonomy (ATO).** [http://purl.bioontology.org/ontology/ATO]
39. **Fly taxonomy (FBsp).** [http://purl.bioontology.org/ontology/FB-SP]
40. **Teleost taxonomy (TTO).** [http://purl.bioontology.org/ontology/TTO]
41. **NCBI organismal classification (NCBITaxon).** [http://purl.bioontology.org/ontology/NCBITaxon]
42. Viljanen K, Tuominen J, Hyvönen E: **Ontology libraries for production use: the finnish ontology library service ONKI.** In *Proceedings of the 6th European Semantic Web Conference (ESWC): May 31–June 4 2009, Heraklion, Greece*. Edited by Aroyo L, Traverso P, Ciravegna F, Cimiano P, Heath T, Hyvönen E, Mizoguchi R, Oren E, Sabou M, Simperl E. Berlin Heidelberg: Springer–Verlag; 2009:781–795.
43. **OBO flat file format 1.4 syntax and semantics [WORKING DRAFT].** 2011. [http://purl.obolibrary.org/obo/oboformat/spec.html]. [Mungall C, Ruttenberg A, Horrocks I, Osumi-Sutherland D (editors)].
44. **TaxonConcept.org.** [http://www.taxonconcept.org]
45. Tuominen J, Laurenne N, Hyvönen E: **Biological names and taxonomies on the semantic web – managing the change in scientific conception.** In *Proceedings of the 8th Extended Semantic Web Conference (ESWC): May 29–June 2 2011; Heraklion, Greece*. Edited by Antonio G, Grobelnik M, Simperl E, Parsia BD, Plexousakis D, Leenheer P, de Pan JZ. Berlin Heidelberg: Springer–Verlag; 2011:255–269.
46. Tuominen J, Laurenne N: **Taxonomic meta-ontology TaxMeOn specification.** 2013. [http://schema.onki.fi/taxmeon]
47. Kirsten T, Gross A, Hartung M, Rahm E: **GOMMA: a component-based infrastructure for managing and analyzing life science ontologies and their evolution.** *J Biomed Semantics* 2011, **2:**6.
48. Maynard D, Peters W, d' Aquin M, Sabou M: **Change Management for Metadata Evolution.** In *Proceedings of the International Workshop on Ontology Dynamics (IWOD), the 4th European Semantic Web Conference (ESWC): 7 June 2007, Innsbruck, Austria*. Edited by Flouris G, d' Aquin M; 2007:27–40.

49. Euzenat J, Shvaiko P: *Ontology Matching*. Berlin Heidelberg: Springer–Verlag; 2007.
50. Khattak AM, Latif K, Lee S: **Change management in evolving web ontologies.** *J Knowledge-Based Syst* 2013, **37:**1–18.
51. Wang S, Schlobach S, Klein M: **Concept drift and how to identify it.** *J Web Semantics* 2011, **9**(3):247–265.
52. **Taxonomic database.** [http://taxon.luomus.fi]
53. DCMI Usage Board: **DCMI metadata terms.** 2012. [http://www.dublincore.org/documents/dcmi-terms/]
54. Beckett D, Berners-Lee T: **Turtle – Terse RDF triple language.** 2011. [http://www.w3.org/TeamSubmission/turtle/]
55. McNeill J, Barrie FR, Buck WR, Demoulin V, Greuter W, Hawksworth DL, Herendeen PS, Knapp S, Marhold K, Prado J, Prud'homme van Reine WF, Smith GF, Wiersema JH, Turland N: *International Code of Nomenclature for algae, fungi, and plants (Melbourne Code), adopted by the Eighteenth International Botanical Congress Melbourne, Australia, July 2011*. Königstein: Koeltz Scientific Books; 2012. [Regnum Vegetabile].
56. **International Comission on zoological nomenclature (ICZN).** [http://iczn.org]
57. TDWG Biodiversity Information Standards: **TDWG taxon rank LSID ontology.** 2007. [http://rs.tdwg.org/ontology/voc/TaxonRank]
58. Silferberg H: *Enumeratio Coleopterorum Fennoscandiae, Daniae at Baltiae*. Helsinki, Finland: Helsingin Hyönteisvaihtoyhdistys; 1992.
59. Silfverberg H: **Enumeratio renovata Coleopterorum Fennoscandiae, Daniae et Baltiae.** *Sahlbergia* 2011, **16**(2):1–144.
60. Kurki J, Hyvönen E: **Collaborative Metadata editor integrated with ontology services and faceted portals.** In *Proceedings of the 1st Workshop on Ontology Repositories and Editors for the Semantic Web (ORES) the 7th Extended Semantic Web Conference (ESWC): 31 May 2010, Heraklion, Greece*. Edited by d'Aquin M, García Castro A, Lange C, Viljanen K: CEUR Workshop Proceedings; 2010:7–11.
61. **Protégé ontology editor.** [http://protege.stanford.edu]
62. **Finnish ontology library service ONKI.** [http://onki.fi]
63. **The environmental observation web and its service applications within the future internet (ENVIROFI).** [http://www.envirofi.eu]
64. Kennedy J, Kukla R, Paterson T: **Scientific names are ambiguous as identifiers for biological taxa: their context and definition are required for accurate data integration.** In *Proceedings of the 2nd International Conference on Data Integration in the Life Sciences (DILS) 20–22 July 2005; San Diego, California*. Edited by Ludäscher B, Raschid L. Berlin Heidelberg: Springer–Verlag; 2005:80–95.
65. **International federation of library associations and institutions (IFLA).** [http://www.ifla.org/]
66. IFLA Study Group on the Functional Requirements for Bibliographic Records: *Functional requirements for bibliographic records : final report*. München, Germany: K.G. Saur; 1998. [UBCIM publications; new series, vol 19].

doi:10.1186/2041-1480-5-40
Cite this article as: Laurenne *et al.*: Making species checklists understandable to machines – a shift from relational databases to ontologies. *Journal of Biomedical Semantics* 2014 **5**:40.

Publication VIII

Jouni Tuominen, Nina Laurenne and Eero Hyvönen. Publishing and Using Plant Names as an Ontology Service. In *Proceedings of the first international Workshop on Semantics for Biodiversity at ESWC 2013*, Montpellier, France, May 27, Pierre Larmande, Elizabeth Arnaud, Isabelle Mougenot, Clement Jonquet, Therese Libourel, Manuel Ruiz (editors), CEUR Workshop Proceedings, volume 979, ISSN 1613-0073, online CEUR-WS.org/Vol-979/WS_s4biodiv2013_paper_2.pdf, May 2013.

Publishing and Using Plant Names as an Ontology Service

Jouni Tuominen, Nina Laurenne, and Eero Hyvönen

Semantic Computing Research Group (SeCo)
Aalto University School of Science, Dept. of Media Technology, and
University of Helsinki, Dept. of Computer Science
http://www.seco.tkk.fi, firstname.lastname@aalto.fi

Abstract. Animals and plants are referred to using scientific or common names depending on the expertise of an audience or a source of data. The names change in time and therefore their usage as identifiers as such is problematic. We present a solution for managing and using plant names as an ontology. The ontology is based on the TaxMeOn meta-ontology for biological names. In order to refer to organisms unambiguously and publish information as Linked Data on the web, the names are given URIs. The ontology is developed collaboratively and it supports the approval process and temporal tracking of the common names. We introduce an ontology service of plant names for end-users and provide user interfaces and APIs for integrating the ontology into applications.

1 Introduction

The scientific names of plants and animals have a major role when indexing, querying, and integrating information about species. Biologists use scientific names although the vast majority of people use the common name equivalents. Contrary to common belief, neither the scientific nor common names identify organisms unambiguously as one name may point to multiple species and one species may have multiple names.

New research results change the name combination of the scientific names because taxa are constantly split and lumped. For example, if a species is changed into another genus, the name combination changes accordingly. Approximately 25,000 new species descriptions are published in thousands of journals annually [6] which makes it hard for researchers to keep up-to-date the biodiversity of the nature. Not all organisms need a common name but still there is huge work to be done in developing the vernacular nomenclature and in terms of established names, the dialect expressions remarkably expand the spectrum of the biological names.

The international commissions of the nomenclatures (IBC, ICZN) specify the rules how the scientific names should be used in various taxonomic treatments. The nomenclatures of plants and animals are independent of each other and the rules are applied only to the scientific names. The common names are not

regulated but they also change in time because there is often a need to update the common names at intervals. The changing nature of the names poses challenges for their management [5, 10, 13].

The diversity of the names causes problems when combining data from heterogeneous sources, e.g., observational records, literature and museum collections [11, 9]. The data cannot be easily integrated if a taxon is referred to using multiple names and vice versa the existence of homonyms (the same name refers to multiple taxa) causes errors when merging the data.

Comprehensive reference lists and catalogues of the names have been proposed as a solution to facilitate the access to the names [1, 10]. However, this is not enough because the biological names ought to be machine-processable in order to refer to them unambiguously and semantically enrich the biological contents. Ontologies remarkably increase the re-use and utilization of the available resources which minimizes the amount of manual work when harmonizing data.

We present an ontology model for managing the common names of organisms and linking them to the scientific names. The model supports temporal tracking of name changes and an approval process of the common names. The model is used for maintaining and publishing plant names in Finnish as an ontology. The ontology is published as Linked Open Data [3] and can be used as an ontology service.

2 Ontology Model

TaxMeOn[1] [14] is an RDF-based meta-ontology for modeling and managing biological names and classifications. TaxMeOn introduces classes and properties for expressing biological names as ontologies. The model consists of three parts according to the level of taxonomic details, which are common names, species checklists, and detailed taxonomic information respectively. In this paper, the focus is on the common names although many of the classes and properties are common to all three parts. The simplified structure of the model is presented in Fig. 1, where the core classes are *Scientific name*, *Common name* and their statuses. The status of the *Scientific name* indicates if a name is an accepted or a synonymized one, etc. The synonyms are linked to an accepted name. The hierarchical structure is constructed setting relations between the *Scientific names*.

The *Common names* (in one or more languages) that refer to the same taxon are connected through a *Scientific name*. The model also allows mapping the scientific names to each other based on the underlying taxonomic concepts (congruence, overlap, part-of, general association). A taxonomic rank expresses the hierarchical level in a classification (e.g., a species, a genus) and it is specified for every scientific name. The taxonomic ranks are presented as a separate vocabulary which contains 61 ranks, of which 60 are obtained from TDWG Taxon Rank LSID Ontology[2]. In order to avoid the complex details of the botanical and

[1] `http://schema.onki.fi/taxmeon/`
[2] `http://rs.tdwg.org/ontology/voc/TaxonRank`

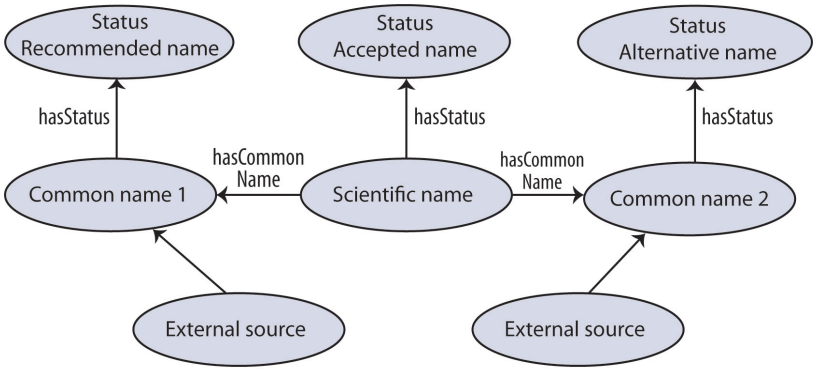


Fig. 1. The ontology model of the common names of organisms. The ellipses represent classes and the arrows depict relations between the classes.

horticultural nomenclatures, the species level and the taxonomic levels below it are treated as one unit.

The approval process of the common names is the following: first, a new name is proposed; then the name becomes accepted; and finally, the name may become an alternative, if there is a new accepted name for the same plant. The model allows the maintainers to propose a new name which then can be commented by the other maintainers until the name finally gets accepted, rejected or synonymized. The temporal management of the names is based on time stamps which are given to the statuses of the names in the approval process. If a name is given a new status, the old status is not removed from the system. This makes it possible to track the chain of changes of the names and to see the period of time period when a particular name was accepted.

3 Managing Plant Names as an Ontology

We applied the TaxMeOn ontology model to a database of the Finnish names of plants maintained by the Finnish Biology Society Vanamo[3]. The original database contained nearly 26,000 plant names in Finnish in a single classification. The taxa were divided into three taxonomic levels (a species, a genus and a family) but it is possible to specify more taxonomic levels in the current ontology.

The database of the plant names was converted into RDF format based on the TaxMeOn ontology model. The ontology is managed in the metadata editor SAHA[4] [7] by the Vanamo association. Currently, the ontology contains 21,797 species, and the number of updates exceeds one thousand names yearly. The

[3] `http://www.vanamo.fi`
[4] `http://www.seco.tkk.fi/services/saha/`

utilization of the ontology facilitates the management of the names because the approval process is integrated into the ontology.

The association has an active role in developing new Finnish names for plants and the public availability of the ontology releases voluntary based work for more relevant activities than responding to various queries by journalists, translators etc. The development of the new names is based on the needs, therefore the coverage of the taxa is not systematically or geographically restricted into any particular plant group or a region.

The browser-based SAHA editor allows collaborative editing of the ontology, providing the simultaneous access of multiple users and a chat functionality. The TaxMeOn model has been extended to support the management of the ontology in SAHA, by adding a property indicating the current status of the processing of a proposed common name. If a new name is suggested for a species, a maintainer can add it into the ontology and mark it as "in process". The proposed but not yet processed names can be found easily at later stages of the process.

4 Using Plant Names as an Ontology Service

The ontology is published as Linked Open Data in the Finnish Ontology Library Service ONKI[5] [15], as part of the Finnish semantic web infrastructure project FinnONTO[6] [4]. The ONKI service provides user interfaces and APIs for accessing and using the plant names in applications. For example, end-users can browse and search the ontology to find a common name for a taxon that they know only by the scientific name. The ONKI selector widget can be integrated into legacy CMS systems to provide an autocomplete and URI fetching features to support the annotation of plant related information.

One of the advantages of the ontology service is that the end-users can now access the ontology themselves. Users are directed to the ONKI service via search engines, and they have adopted the service by extending Wikipedia articles about plant species with links to Finnish plant names in ONKI. End-users actively send feedback, comments and corrections to the maintainers, which help them to improve the quality of the content.

The ontology is also accessible as a SPARQL endpoint. An example query below shows how the accepted Finnish common names of species (and taxa below it) that belong to a genus "*Quercus*" (oak) can be retrieved:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX taxmeon: <http://www.yso.fi/onto/taxmeon/>
PREFIX taxonomic-ranks: <http://www.yso.fi/onto/taxonomic-ranks/>

SELECT ?vernacularName WHERE {
  ?species taxmeon:isPartOfHigherTaxon ?genus .
  ?genus rdf:type taxonomic-ranks:Genus .
  ?genus rdfs:label "Quercus"^^xsd:string .
```

[5] http://onki.fi/en/browser/overview/kassu

[6] http://www.seco.tkk.fi/projects/finnonto/

```
  ?species taxmeon:hasVernacularName ?vernacularNameRes .
  ?vernacularNameRes taxmeon:hasVernacularNameStatus ?status .
  ?status rdf:type taxmeon:AcceptedVernacularName .
  ?vernacularNameRes rdfs:label ?vernacularName .
  FILTER langMatches(lang(?vernacularName), "fi")
}
```

The result of the query is a list of the Finnish names of oak species, such as the sessile oak and white oak. The query demonstrates the use of the ontology for cross-language query expansion.

Currently, the plant name ontology is used by several cultural museums and libraries for annotating their collections. The ontology is also applied as a use case in the EU FP project ENVIROFI[7] which focuses on the environmental usage area of the Future Internet. The ontology is used in the project as a conceptual hub for referring to the plants in the observational data on biodiversity. The ontology has been extended with the English and German names of plants used in the project pilots (these names are not available in the ONKI ontology service).

5 Discussion

5.1 Related Work

The importance of persistent identifiers for organism names and solutions for managing them on the semantic web have been discussed by several workers. Page [8] presented how taxon names are modeled as semantic metadata in RDF form. Taxon names are identified with using Life Science Identifiers (LSID) and the names are connected using taxonomic relations. Taxon names that are obtained from various data sources and which refer to the same taxon are mapped using the *owl:sameAs* relation. Schulz et al. [12] presented the first ontology model of biological taxa and its application to physical individuals. The model is based on a single unchangeable classification. Franz and Thau [2] evaluated the limitations of applying ontologies to the scientific names and concluded that ontologies should focus either on a nomenclatural point of view or on strategies for aligning multiple taxonomies.

The Darwin Core (DwC)[8] is a metadata schema developed for taxon occurrence data by the TDWG (Biodiversity Information Standards). The goal of DwC is to standardize the form of how biological information is presented. However, it lacks the semantic aspect and when it comes to the names, the scope of DwC is quite general.

Taxonconcept.org[9] provides Linked Open Data identifiers for species concepts and links data from different sources. All the names of species are expressed using literals. Also, the machine-processability is weakened by the usage of literal values for expressing the hierarchies. The data contains scientific and common names, and taxonomic statuses.

[7] http://www.envirofi.eu
[8] http://www.tdwg.org/standards/450/
[9] http://www.taxonconcept.org

Many existing databases aim to be comprehensive online catalogues that aggregate individual species checklists, such as the Catalogue of Life (CoL)[10] and The International Plant Names Index (IPNI)[11]. The IPNI database contains only scientific names, but the Catalogue of Life also includes their taxonomic statuses and common names. They both provide the names in a machine-processable form, as RDF conforming to the TDWG Taxonomic Concept Transfer Schema (TCS)[12] using LSIDs as identifiers of the names [5]. In the Catalogue of Life the requirement to use a separate LSID resolver for fetching metadata about an LSID prevents the Linked Data compatibility of the dataset. The IPNI database provides an LSID proxy that allows Linked Data compatibility. In the IPNI database, the hierarchy is not expressed explicitly in the RDF (e.g., the genus of a species is shown only in the binomial name literal).

There are several other plant name databases available on the web, e.g., the Royal Horticultural Society Horticultural Database[13], The Plant List[14] and the Euro+Med PlantBase[15]. Most available resources contain the scientific names, but in few, the common names are included. Common to these systems is that they are intended for human usage, and they are not available in a machine-processable form with unique name identifiers.

5.2 Contributions and Future Work

Most of the related work concentrate on the scientific names, but our focus is on the common names. The common names expand the cross-domain use of the ontology because they are in wider spectrum of use than the scientific ones. The ontology is available in machine-processable RDF format, with explicit semantics, e.g., the hierarchical relations are set between the plant URIs, and the statuses of names are supported. The TaxMeOn model provides a solution for managing the approval process of common names, supporting the temporal tracking of the name changes via statuses and their time stamps. The model connects together different names of a taxon facilitating data integration and information retrieval in cases where data is combined from heterogeneous sources.

We have also demonstrated the complete workflow from a collaborative development of an ontology to publishing it as Linked Open Data and as an ontology service which makes it accessible to the general public. The plant name ontology helps harmonizing the terminology which in turn enhances communication between various users. Application developers can utilize the ontology by using the plant name URIs for unambiguous referencing to plants species.

Currently, hybrid taxa are modeled in the ontology in the same way as the ordinary species. An idea for the future development is to extend the model to

[10] `http://www.catalogueoflife.org`
[11] `http://www.ipni.org`
[12] `http://www.tdwg.org/standards/117/`
[13] `http://apps.rhs.org.uk/horticulturaldatabase`
[14] `http://www.theplantlist.org`
[15] `http://www.emplantbase.org`

support the representation of hybrid names at a deeper level. Another area for development is to link the scientific names of plants to their author URIs in DBpedia, connecting the ontology to the Linked Data Cloud (LOD).

Ontologies are a bridge between experts and ordinary people in communication and popularizing science. Additionally, the Linked Data approach provides a way how to easily extend an ontology with additional information which in turn increases the information value of contents.

Acknowledgments This work is part of the National Semantic Web Ontology project in Finland FinnONTO (2003-2012), funded mainly by the National Technology and Innovation Agency (Tekes) and a consortium of 38 public organizations and companies, and the EU FP project The Environmental Observation Web and its Service Applications within the Future Internet (ENVIROFI). We thank Leo Junikka and Arto Kurtto for their collaboration.

References

1. Dengler, J., Berendsohn, W.G., Bergmeier, E., Chytrý, M., Danihelka, J., Jansen, F., Kusber, W.H., Landucci, F., Müller, A., Panfili, E., Schaminée, J.H.J., Venanzoni, R., von Raab-Straube, E.: The need for and the requirements of EuroSL, an electronic taxonomic reference list of all european plants. Biodiversity & Ecology 4, 15–24 (2012)
2. Franz, N., Thau, D.: Biological taxonomy and ontology development: scope and limitations. Biodiversity Informatics 7, 45–66 (2010)
3. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1–136, Morgan & Claypool (2011)
4. Hyvönen, E., Viljanen, K., Tuominen, J., Seppälä, K.: Building a national semantic web ontology and ontology service infrastructure—the FinnONTO approach. In: Proceedings of the ESWC 2008, Tenerife, Spain. Springer–Verlag (2008)
5. Jones, A.C., White, R.J., Orme, E.R.: Identifying and relating biological concepts in the Catalogue of Life. Biomedical Semantics 2(7) (2011)
6. Knapp, S., Polaszek, A., Watson, M.: Spreading the word. Nature 446, 261–262 (2007)
7. Kurki, J., Hyvönen, E.: Collaborative metadata editor integrated with ontology services and faceted portals. In: Workshop on Ontology Repositories and Editors for the Semantic Web (ORES 2010), the Extended Semantic Web Conference ESWC 2010, Heraklion, Greece. CEUR Workshop Proceedings, `http://ceur-ws.org` (2010)
8. Page, R.: Taxonomic names, metadata, and the semantic web. Biodiversity Informatics 3, 1–15 (2006)
9. Page, R.D.M.: Biodiversity informatics: the challenge of linking data and the role of shared identifiers. Briefings in Bioinformatics 9(5), 345–354 (2008)
10. Patterson, D.J., Cooper, J., Kirk, P.M., Pyle, R.L., Remsen, D.P.: Names are key to the big new biology. Trends in Ecology & Evolution 25(12), 686–691 (2010)
11. Sarkar, I.N.: Biodiversity informatics: organizing and linking information across the spectrum of life. Briefings in Bioinformatics 8(5), 347–357 (2007)

12. Schulz, S., Stenzhorn, H., Boeker, M.: The ontology of biological taxa. Bioinformatics 24(13), 313–321 (2008)
13. Segers, H., de Smet, W.H., Fischer, C., Fontaneto, D., Michaloudi, E., Wallace, R.L., Jersabek, C.D.: Towards a list of available names in zoology, partim phylum rotifera. Zootaxa 3179, 61–68 (2012)
14. Tuominen, J., Laurenne, N., Hyvönen, E.: Biological names and taxonomies on the semantic web – managing the change in scientific conception. In: Proceedings of the ESWC 2011, Heraklion, Greece. pp. 255–269. Springer–Verlag (2011)
15. Viljanen, K., Tuominen, J., Hyvönen, E.: Ontology libraries for production use: The Finnish ontology library service ONKI. In: Proceedings of the ESWC 2009, Heraklion, Greece. pp. 781–795. Springer–Verlag (2009)

Knowledge organization systems such as thesauri and ontologies are used for improving the findability of information. They harmonize the content descriptions and provide interoperability in and between information systems, such as museum collections, digital libraries, and online stores.

Ontology services facilitate the use of knowledge organization systems in applications. This thesis presents methods for cost-effective ontology access for content indexers, information searchers, and application developers.

The developed tools are provided as a public living lab service, and they facilitate the simultaneous use of ontologies from different domains, such as music, economics, and agriculture. A method is presented for managing the changes of biological taxonomies as an ontology.

Cover image:
The concept hierarchy of the top three levels of the KOKO ontology

ISBN 978-952-60-7456-6 (printed)
ISBN 978-952-60-7455-9 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Computer Science
www.aalto.fi

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL DISSERTATIONS

Aalto-DD 101/2017