BookSampo—Lessons Learned in Creating a Semantic Portal for Fiction Literature

Eetu Mäkelä[1], Kaisa Hypén[2], and Eero Hyvönen[1]

[1] Semantic Computing Research Group (SeCo),
Aalto University and University of Helsinki, Finland
first.last@aalto.fi, http://www.seco.tkk.fi/
[2] Turku City Library, Turku, Finland
first.last@turku.fi

Abstract. BookSampo is a semantic portal in use, covering metadata about practically all Finnish fiction literature of Finnish public libraries on a work level. The system introduces a variety of semantic web novelties deployed into practise: The underlying data model is based on the emerging functional, content-centered metadata indexing paradigm using RDF. Linked Data principles are used for mapping the metadata with tens of interlinked ontologies in the national FinnONTO ontology infrastructure. The contents are also linked with the large Linked Data repository of related cultural heritage content of CultureSampo. BookSampo is actually based on using CultureSampo as a semantic web service, demonstrating the idea of re-using semantic content from multiple perspectives without the need for modifications. Most of the content has been transformed automatically from existing databases, with the help of ontologies derived from thesauri in use in Finland, but in addition tens of volunteered librarians have participated in a Web 2.0 fashion in annotating and correcting the metadata, especially regarding older litarature. For this purpose, semantic web editing tools and public ONKI ontology services were created and used. The paper focuses on lessons learned in the process of creating the semantic web basis of BookSampo.

1 Introduction

With the advent of the Internet, the role of libraries as primary sources of factual knowledge has diminished, particularly among younger people. This is reflected in many of the analyses published in a 2011 study of library use in Finland [13]. Even taken as a whole, 83% of the respondents of the study said they rather sought factual knowledge from the Internet than from the other tallied channels of public libraries, television, magazines or friends. Even for deeper learning on a subject, 40% favored the Internet, as opposed to 38% who still preferred the library. At the same time, the role of the library as a provider of fiction literature has remained constant. While 34% of the respondents said they still benefited from factual information stored in libraries, the corresponding percentage for fiction was 45%.

These results encourage libraries to improve their services related to fiction literature, such as the search and recommendation facilities available for such content. However, the special nature of fiction necessitates a move from the old library indexing traditions, i.e. mainly classifying books by genre and by cataloguing their physical location, to describing their content. This is a very recent development in the centuries long timescale of libraries,

In Finland, content keywords for fiction have been systematically entered only since 1997, using the Finnish fiction content thesaurus Kaunokki[3], developed since 1993. The reason for this is twofold. First, there is a long running tradition in libraries of favoring assigning a singular classification to each book. This is simply due to the relatively recent advent of library computer systems, starting in Finland in the 1980s. Before, when book information was stored on physical index cards, any added keywords past the necessary single classification necessitated also adding another physical card to the indexing cabinets. For fiction, this has always been somewhat problematic, as it appeared quite impossible to arrive at a single universal best classification system, even though various attempts at such have been proposed from as early as the 1930s [11].

Even after the single classification issue was resolved, fiction content description was still considered almost impossible due to the interpretational ambiguity of texts. It was also feared that content description with only a few words would abridge the connotations of a work, and could actually do more harm than good to literature and its multidimensionality. Experiments in indexing consistency from 1999 however found that there was much uniformity in how individual readers paid attention to certain fictional works, and that most fictional content could be adequately described by about 10 to 15 indexing terms [11].

On the other hand, the study also concluded that customers descriptions of the pertinent points of, and questions about fiction literature tended to combine details related to for example the author, contents and publication history of a given work. Based on this, the author of the study compiled a wide or ideal model for describing fiction, which in his mind should include not only the factual publication data on the book, but also descriptions of the content, information on any intertextual references contained therein and data about the author, as well as information about the reception of the book by readers at different times, interpretations by researchers and other connections that help position the publication in its cultural historical context.

In 1999, this model was considered an ideal, impossible to implement in reality. However, times change, and when the Finnish public libraries in the summer of 2008 started a joint venture to experiment with new ways of describing fiction, the model was chosen as a concrete goal to strive for. At this point, based on knowledge gained from prior collaboration, the libraries approached the Semantic Computing Research Group at the Aalto University and University of Helsinki, and the BookSampo project started as part of the national FinnONTO initiative[4]. Because the model that was strived for placed much emphasis on the

[3] http://kaunokki.kirjastot.fi/

[4] http://www.seco.tkk.fi/projects/finnonto/

interconnections between heterogeneous information items, it seemed a good fit for semantic web technologies. For example, the cultural historical context of fiction literature is not restricted to other literature, but spans the whole of culture and history from the time of their writing onwards. Books are nowadays also often sources for movies etc., further demanding a broadening of horizons.

The research group had also much prior experience in publishing cultural heritage content on the semantic web, having created the MuseumFinland portal [6] in 2004 and the CultureSampo portal [5, 8] in 2009. The BookSampo project could make use of much of the infrastructure created for CultureSampo in converting legacy data and thesauri to RDF and ontologies respectively, in editing and annotating data semantically, and in providing intelligent end-user services based on that data.

A beta-version of the end-user portal developed is currently available at `http://www.kirjasampo.fi/`. Already, the portal contains information on virtually all fiction literature published in Finland since the mid 19th century, a total of some 70 000 works of fiction, 25 000 authors and 2 000 publishers.

In the following, lessons learned during the BookSampo project will be presented. First discussed are the many insights gained in modelling fiction, both from an abstract viewpoint as well as how it relates to the semantic web. After that, experiences gained while transforming Kaunokki into an ontology are given. Then the paper presents a technical description of various parts of the system, focusing on the challenges faced and benefits gained from applying semantic web technologies. Finally, the paper discusses the further potential and reception of the developed system in library circles.

2 The BookSampo Data Model

From the start, the BookSampo project was consciously geared towards an ambitious, disruptive experiment as opposed to an incremental improvement. Thus, early on it was decided that the MARC format[5] still currently used in most libraries in Finland would not be used in the project on account of its restrictions, nor would the system be built on top of the current library indexing systems. Instead, the data model of BookSampo was to be based purely on RDF and Semantic Web principles, with both the indexing machinery as well as the public end user interface operating directly on that data to best fulfil its potential.

One of the benefits of the RDF data model is its flexibility. This proved an asset in the project. Because of the experimental nature of the project, there have been multiple times when the model has needed amendment and modification. In addition to simple addition of fields or object types, the schema has undergone two larger alterations during the project.

First, the way the biographical information of the authors was encoded was changed from events to attributes. Initially, details about, among others, the times and places of authors' births, deaths and studies were saved in BookSampo

[5] http://www.loc.gov/marc/

as events, in the spirit of the cultural heritage interchange model of CIDOC-CRM [2] and the BIO-schema of biographical information [1].

User research, as well as interviewing library indexers revealed, however, that events as primary content objects are not easily understood by those indexing them or by end-users on a cognitive level. Bringing events to the fore, the approach fractured and distributed the metadata of the original primary objects. For example, people wanted much more to see information on authors' birth and death dates and places as simply attribute-object values of the author, instead of as events where the author was involved in.

Description thus adopted back the more traditional model, where data about times and places of occurrences are directly saved as author attributes. In the case of studies, this did lead to some loss of data specificity, as the original information related for example the dates and places to each individual degree attained. This information could not be maintained in a flat attribute value description centered on the author. However, the indexers deemed the simplicity to outweigh the costs in this situation.

An even larger change however was made to the book schema. It has been a conscious policy that BookSampo should only concentrate on the description and data concerning the contents of the work itself, irrespective of editions. But right from the start, details about translators, publication years, publishers and publishing series crept in. The guidelines at the time were to save only the details of the first Finnish edition.

For a very long time, the model of a single object worked well, until it was decided that the project should also extend to include Swedish literature[6], as well as maintain distinctions between different translations. It then became necessary to reconsider how the different conceptual levels of a work could be separated from each other. Advice was sought from the FRBRoo Model [10], which identifies four conceptual levels, among which the different properties of a work can be divided:

1. Work. The abstract contents of the work—the platonic idea of the work (primary creator, keywords).
2. Expression. The concrete contents of the work — original/translated text, stage script and film script (author, translator and director).
3. Manifestation. The concrete work/product—book, compilation book, entire concept of a stage performance and film DVD (publisher, issuer and ISBN).
4. Item. The physical copy—single book/compilation/performance/DVD.

The idea in the model is that a single abstract conceptual work may be written down in different texts, which may then be published in multiple editions, each comprised of a multitude of physical copies. Each type of item down the scale inherits the properties given on the levels above it. Translated into actual

[6] Finland is a bilingual country, the official languages being Finnish and Swedish. This is why a web service maintained by the public libraries must be available in both languages.

indexing work, this means that for example the content of a work need be described only once, with each different language edition merely referring to the resource describing the qualities of the abstract works printed therein.

After what had been learnt from the biography schema, it was not deemed desirable to replace a simple model with the complexity of four entire levels. Also, more generally, experience had proven that the BookSampo indexing model focusing on the contents of the work was already quite a leap to librarians, who were thus far mostly familiar with single level MARC indexing of mostly manifestation level information.

Since data in BookSampo never reaches the level of a single item, it was easy to simply drop the item level. On the other hand, the work level had to be kept separate, so translations in different languages could refer to the same content keywords. It was decided, however, to combine the expression and manifestation levels, since, one translation has on the average one publisher and physical form, and repetitive descriptions would hence not be needed on a large scale.

As a result, works are described at two levels in BookSampo: as an abstract work, which refers to the contents of the work, which is the same in all translations and versions and as a physical work, which describes features inherent to each translation or version.

The following fields are used to describe an abstract work:

- work titles in all the languages in which the physical work is stored
- author(s)
- type (novel, novella, poem, etc.)
- genre (fantasy literature, one-day novel, detective novel, etc.)
- subjects and themes addressed in the work
- actors (in general, children, doctors and middle-aged people)
- main characters (proper names, e.g. Adam Dalgliesh)
- time of events (in general, summer, middle ages or a specific dating)
- place of events (in general, libraries, countryside or a real-world place)
- textual summary or presentation
- language of the original work
- work adaptations (movies, operas, plays, etc.)
- related works (librarian recommendations)
- information related to the work on the web (critiques, descriptions, etc.)
- awards

The following fields are related to the description of a physical work:

- title and subtitle
- original title
- language of the work
- publisher
- date published
- number of pages
- translator
- illustrator
- editor or other contributor
- additional information on publication history (edition, illustrations, etc.)
- serial information

In addition, a book cover image can be included with a physical work, and it is also possible to describe the cover with keywords, provide a brief description of it and list the cover designer. It is also possible to associate review objects to the abstract works, describing for example contemporary reader reactions.

The physical works are linked with the abstract work and their data are presented in the context of the abstract work. This way it is possible to demonstrate, for example, that Hopealaiva and Nostromo are both Finnish translations of Nostromo, a Tale of the Seaboard by Joseph Conrad.

While it can be argued that not using the whole FRBR model diminishes the interoperability of the content in BookSampo with regard to other FRBR collections, it turns out that also others have independently come to a similar simplification of the model, particularly in systems where distributed editing and understandability of content is important, as opposed to for example systems doing automatic conversion of MARC records to FRBR. For example, the Open library[7] recognizes work and edition levels, with the latter also combining expression and manifestation. Exactly the same situation is present also in the LibraryThing portal[8], only naming the entities as “work” and “book”. On the other hand, even systems that claim to support separate expression level items on their data model level, such as The Australian Music Centre[9], and the OCLC WorldCat system[10], do not allow these to be shown or searched for independently of their linked work or manifestation entities in their actual user interfaces, thus further corroborating that at least from an end-user perspective, the distinction between an expression and a manifestation is not very important.

In any case, it has already been established by others that separation of expressions from even the original combined MARC fields is possible by mostly automated processing along with manual correction [3, 9], so should a need for further separation arise, one could just repeat a similar split procedure as just presented for the BookSampo data.

In BookSampo, the experience of moving from the solution of one conceptual level to that of two was mainly simple and painless. A minor problem was, however, short stories and their relationship with short story collections. Originally, two objects here were turned into four, and their internal relationships required precise inspection. Finally, it was decided to choose a model where each short story had an abstract work level, which was “embodied” as a “part of a physical work”. This “part of a physical work” was then included in a physical work, which in turn was the “embodiment” of the short story collection as an abstract work. This set of relationships is depicted in a more visual form in figure 1.

This way both the individual short story and the short story collection overall may separately have content keywords. Whereas most of the data at the manifestation level belongs to the physical work of the short story collection, the data of an individual short story at the expression level, e.g. details of the translator, the name in the collection or page numbers, belongs to the part of the physical

[7] http://www.openlibrary.org/

[8] http://www.librarything.com/

[9] http://www.australianmusiccentre.com.au/about/websitedevelopment

[10] http://frbr.oclc.org/pages/

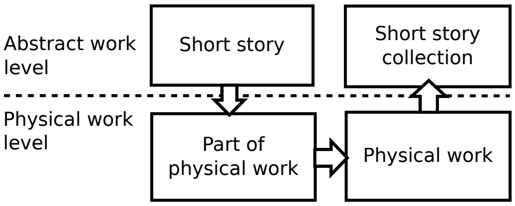


Fig. 1. Relationship between short story and short story collection in the BookSampo data model.

work. This same structure is also applied to other similar part-object instances, for example single poems in poem collections.

As a benefit of the flexibility of the RDF data model, all the transformations described here could be done by quite simple transformation scripts that operated only on the data, without having to change editor or database code.

There is one area where the RDF data model caused problems. For the most part, the model is simple. Each resource describes an independently existing thing, such as a book, award, author or a place, that has relationships with other things. Yet even this was sometimes hard for people who were used to each annotation being completely separate. For example, at one point it was discovered that two different URIs for the 1970s had crept into the database. Upon closer inspection, it was discovered that one of them was actually the URI for 1960s, only someone had changed the label when they were trying to correct a particular book's annotation as happening in the 1970s instead of the 1960s.

However, a much greater problem was the confusion arising from cases where a particular resource actually didn't note an independently existing, understandable thing. This has to do with cases where a relation has metadata of its own, such as when one wants to record the year a book has been given an award or the serial number of a book in a series. In RDF, these situations are usually resolved by creating the link through an anonymous node where this information can be recorded. For example, a book can be related to the resource "Part 7 in the Yellow Library", which is in turn annotated as a part of the Yellow Library series with a part number of 7.

In BookSampo, this caused problems because these auxiliary resources appeared to the user exactly like resources describing something independently extant, yet their function was different—i.e. it doesn't really make sense to think that "Part 7 of the Yellow Library series" exists in any sense separate from the book that holds that place in the series. In BookSampo, there was no way around using these auxiliary resources for certain things, but their use certainly did muddy the primary concept of the graph model. Luckily, in practice the effects of this could be somewhat mitigated by training and annotation instructions. However, further developments in generalized visualization, browsing and editor environments should do well to provide support for special handling of such auxiliary resources, so that such inconsistencies arising from technical limitations can be hidden behind better user interfaces.

3 Ontology Creation and Participation in the FinnONTO Infrastructure

To maximally leverage the possibilities of semantic web technologies in the project, the Finnish and Swedish thesauri for fiction indexing (Kaunokki and Bella) were converted into a bilingual ontology KAUNO [12]. This was done in order to allow the use of inferencing in the recommendation and search functionalities of the end-user portal.

This ontology was also linked with other available ontologies and vocabularies used in CultureSampo, to improve the semantic linking of the content to their cultural historical context. Here, the project leveraged and relied on the work done in the wider FinnONTO project [4], which aims to make uptake of the semantic web as cost-effective as possible in Finland by creating a national infrastructure for it. At the core of the FinnONTO model is the creation of a national ontology infrastructure termed KOKO, which aims to join and link together under an upper ontology as many domain specific ontologies as possible.

The primary core of KOKO is currently comprised of 14 domain ontologies joined under the the Finnish national upper ontology YSO[11]. Among these are for example the museum domain ontology MAO, the applied arts ontology TAO, the music ontology MUSO, the photography domain ontology VALO and the ontology for literature research KITO. Thus, by linking the fiction literature ontology KAUNO to KOKO, links would immediately be created bridging the material to all this other cultural heritage.

Thus far, with the exception of MAO and TAO, these ontologies are each joined only through common links to YSO. This has been possible because almost all common concepts of the domain specific are also in YSO, and domain concepts appear mostly only in that domain. This was the approach taken also with regard to KAUNO. To create the links, automatic equivalency statements between the concepts of the KAUNO and YSO ontologies were generated by tools created in the FinnONTO project. After this, the combined ontology file was loaded into the Protégé ontology editor[12]. All automatically created links were checked by hand, as well as new links created.

The experience of the librarians who ontologized Kaunokki was that it brought in a very welcome additional structuring to the vocabulary. For example, the term "american dream", in the thesaurus only contained information that it belonged to the theme facet. In the ontology however, it had to find a place in the ontology's class hierarchy: a lifestyle, which in turn is a social phenomena, which at the root level is an enduring concept (as opposed to a perduring or abstract concept). This forced additional work ensures that no keyword floats around in isolation, but is always surrounded by co-ordinate, broader and narrower concepts that help define it and relate it to other phenomena. This also beneficially forces the vocabulary keeper to narrow down and elucidate their definition of the keyword, which in turn helps in ensuring uniform use of keywords by indexers.

[11] http://www.yso.fi/onki3/en/overview/yso

[12] http://protege.stanford.edu/

The linking to YSO was also deemed extremely beneficial, as before, even all general keywords were maintained in each vocabulary separately. Now, their management could be centralized, while having the work done still be usable as part of systems utilizing the domain ontologies.

4 System Description

In this section, the paper presents the technical solutions created for the BookSampo System, focusing on the challenges faced and benefits gained from applying semantic web technologies. First, the data import and integration functionality used to both bootstrap as well as update the system is discussed. Then presented is the primary editing environment created for the dozens of volunteers distributed in Finnish libraries who do content enrichment for the project. Finally, the functionality used to built the end-user portal search and browsing functionality is discussed.

4.1 Data Import and Integration

Contrary to existing library systems, the project was not to focus on the characteristics of individual physical editions of books, but equally to the content as well as the context of the different conceptual works themselves. However, it still made sense to bootstrap the database from existing collections, where possible.

The BookSampo project needed first to do data importing, conversion and integration. Data on books was to be sourced primarily from the HelMet cataloguing system[13] used in the Helsinki metropolitan area libraries, which stored its information in the ubiquitous MARC format. Also, from very early on, the vision of the project included storing information not only of the books, but also of the authors of those books. Data on these were to be sourced from three different databases maintained at various county libraries across Finland. Thus, at the very beginning, the project already had at least two quite different content types, and multiple data sources.

The data converters were created following principles established in the CultureSampo project [8], which focus on first converting a data set into RDF as is in an isolated namespace, and then using common tools to map them to each other, as well as to external ontologies and place and actor registries on a best-effort basis. Doing the mapping at the end on the RDF data model and ontology language level, such as using owl:sameAs mappings, brings multiple benefits. First, it ensures that no primary information is lost in the conversion, what often happens when trying to map text-based keywords in the original data to a curated ontology immediately in place. Second, it also allows any automatic mappings to be later reversed if found problematic, as well as iteratively adding more mappings as resources permit. This can be either done manually, or when centrally developed automatic mapping tools improve.

[13] http://www.helmet.fi/search~S9/

In the case of BookSampo, the format of the original author records quite closely matched the end schema sought also in the BookSampo system. However, the book records to be imported were in the edition-centric MARC format. Here, each edition in the source data was simply converted into an abstract work in the BookSampo schema. A large number of volunteers in libraries then poured through the data, manually removing and joining duplicate works that had resulted from multiple editions of a single work in the source.

The conversion of the records from MARC to linked RDF already bought an instant benefit to the project: Before, the fiction content descriptions had been stored in the HelMet library system only as text fields containing the Finnish language versions of the keywords. Now, when they had been converted into URI references in the bilingual ontology, they could instantly be searched using either language. Also, because YSO was available also in English, much of the content could additionally now also be searched in that language. In addition, the use of the CultureSampo authority databases allowed the automatic unification of different forms of author names found in the system, while the place registries of CultureSampo instantly added geo-coordinate information to the place keywords for later use in creating map-based user interfaces to the data.

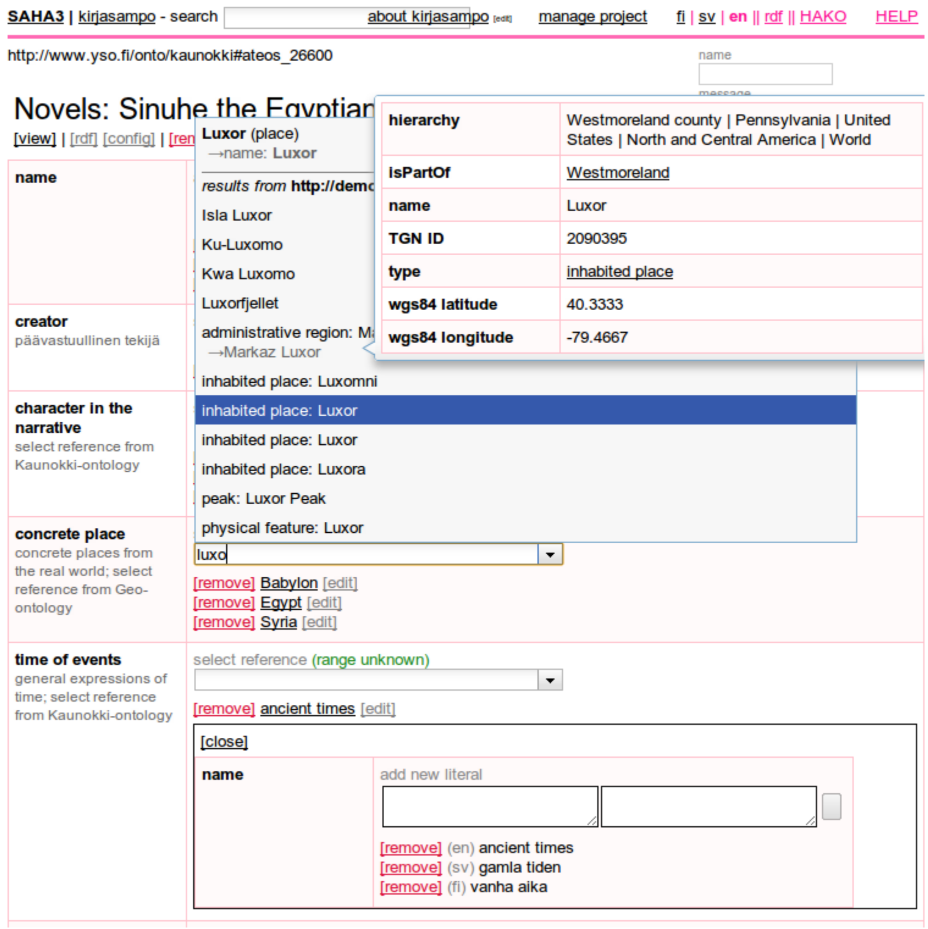
Recently, the BookSampo project also bought rights to descriptions of newly released books from BTJ Finland Ltd, a company that provides these descriptions to Finnish library systems for a price. These descriptions are fetched from the BTJ servers each night in the MarcXML format used also for HelMet, automatically converted to RDF using the CultureSampo tools, and added to the RDF project with tags indicating they should be verified. The librarians then regularly go through all such tagged resource in the editing environment, removing the "unverified" tags as they go along.

4.2 Collaborative Semantic Web Editing Environment

As BookSampo didn't have a data store or editor environment of its own ready, the project decided to adopt the SAHA RDF-based metadata editor [7] developed by the FinnONTO project as its primary editing environment.

SAHA is a general-purpose, adaptable editing environment for RDF data, capable of utilizing external ontology services. It centers on projects, which contain the RDF data of a single endeavour. The main screen of SAHA provides a listing of the class hierarchy defined in the project, from which new instances can be created. Alternatively, existing instances of a particular type can be listed for editing, or a particular instance sought through text search facilities. Navigation from resource to resource is also supported in order to examine the context of a particular resource, with pop-up preview presentations allowing for even quicker inspection of the resources linked.

The editing view of the SAHA editor is depicted in figure 2. Each property configured for the class the resource is an instance of is presented as an editor field, taking in either literal values or object references. For object references, SAHA utilizes semantic autocompletion. When the user tries to find a concept, SAHA uses at the same time web services to fetch concepts from connected external ONKI ontology repositories [15], as well as the local project. Results are shown in one autocompletion result list regardless of origin, and their properties can also be inspected using pop-up preview presentations. In the example depicted in figure 2 for example, this is extremely useful when the user must choose which of the many Luxors returned from both local and external sources is the correct annotation for this book.



SAHA3 | kirjasampo - search about kirjasampo [edit] manage project fi | sv | en || rdf || HAKO HELP

http://www.yso.fi/onto/kaunokki#ateos_26600

name

Novels: Sinuhe the Egyptian

[view] | [rdf] [config] | [ren

name

creator
päävastuullinen tekijä

character in the narrative
select reference from Kaunokki-ontology

concrete place
concrete places from the real world; select reference from Geo-ontology

Luxor (place)
→name: Luxor

results from **http://demo**

Isla Luxor
Ku-Luxomo
Kwa Luxomo
Luxorfjellet
administrative region: M
→Markaz Luxor
inhabited place: Luxomni
inhabited place: Luxor
inhabited place: Luxor
inhabited place: Luxora
peak: Luxor Peak
physical feature: Luxor

hierarchy	Westmoreland county | Pennsylvania | United States | North and Central America | World
isPartOf	Westmoreland
name	Luxor
TGN ID	2090395
type	inhabited place
wgs84 latitude	40.3333
wgs84 longitude	-79.4667

luxo

[remove] Babylon [edit]
[remove] Egypt [edit]
[remove] Syria [edit]

time of events
general expressions of time; select reference from Kaunokki-ontology

select reference (range unknown)

[remove] ancient times [edit]

[close]

name

add new literal

[remove] (en) ancient times
[remove] (sv) gamla tiden
[remove] (fi) vanha aika

Fig. 2. The SAHA metadata editor, showing both semantic autocompletion as well as a pop-up preview presentation of one of the autocompletion results.

For the purposes of the BookSampo project, the SAHA editor was improved with an inline editor feature. The idea is simple: a resource referenced through an object property can be edited inline in a small version of the editor inside the existing editor. Specifically, this functionality was developed to ease the use of the necessary auxiliary resources discussed before. However, there seemed no reason to restrict the functionality to those alone, so this possibility is now available for all linked object resources. In figure 2, this is shown for the property 'time of events" whose value "ancient times" has been opened inline for editing.

From the library indexers point of view, a major source of excitement in the RDF data model and the SAHA editor has been their support for collaborative simultaneous editing of a rich, semantically linked network. Libraries in Finland have shared MARC records between each other for a long time, but these go from a silo to another, and always as whole records focused on individual book editions. In SAHA by contrast, newly added authors or publishers for example, along with all their detailed information are immediately available and usable for all the dozens of voluntary BookSampo indexers across Finland. Once entered, publisher information need also not be repeated again for all new books, which adds an incentive to provide richer detail about also these secondary sources. Similarly, adding a detail to any node in the graph immediately adds value also to all items linked to that node, benefiting information seekers everywhere. In the words of the indexers, this has been both a revelation and a revolution. To further foster co-operation in the SAHA editor between peer indexers, a project-wide chat facility is shown on the top right of each page, facilitating instant discussions (not visible in figure 2 because of occlusion by the pop-up preview).

A similar source of acclaim has been the semantic autocompletion functionality of SAHA. While previously keywords had to be looked up in two different applications separately and copied to the indexing software by hand, or entered from memory leading to typing errors, now they are easily available in a joined list, with the pop-up presentation view allowing for quickly evaluating possible keywords in place. Also valued is the possibility in SAHA for creating new keywords inside the project if no existing keyword suffices. Previously, this would have gone completely against the thought benefits of having a controlled search vocabulary in the first place. However, in SAHA and with ontologies creating e.g. new concepts or locations is not detrimental, provided that the indexer then uses the inline editing functionality of SAHA to link the newly created resource to the existing ontology hierarchies.

All in all, the majority of indexers taking part in BookSampo indexing have found the SAHA editor to be both intuitive, as well as even inspiring. In many cases however, this happened only after an initial confusion caused by the system having both a foreign data model as well as employing a new functional and content indexing paradigm.

4.3 End-User Portal

The complexity of the BookSampo data model entailed problems for the search service side of CultureSampo, which was to be used in the project as the underlying content service. The user interface of BookSampo is being built on top of the Drupal[14] portal system. However, the intention of the project has always been that all primary information be kept and served by CultureSampo, with the Drupal layer only adding commenting and tagging functionality, forums, blogs etc. on the client side.

The problem then became that the CultureSampo search interfaces at the time could only efficiently deal with text queries targeting a single RDF resource, while the BookSampo model required much more complex queries. Particularly, the model required keyword queries to span multiple resources, and combine with relation constraints in the style of [14]. This is because in BookSampo, besides splitting each book into two objects, the model already contained many objects related to the book that needed to be parsed in order produce a search result as well as to render a complete visualization of the book in a user interface. For example, the intent of the BookSampo search interface was that for example the plain text query "Waltari Doctor Inscriptions" would match the abstract work "Sinuhe the Egyptian", because Waltari is the book's author, because it has a main character who is a doctor and because one of its editions has a cover that contains hieroglyphs, which are a type of inscription.

However, inside the data model, this is quite a complex pattern, as visualized in figure 3. First, each resource with a label matching any of the keywords must be found. This results in a result set with (among others) Mika Waltari the author and the keywords doctor and inscriptions. Then, all resources relating to these or their subconcepts must be added to the result set. This results in (among others) the abstract work Sinuhe the Egyptian (whose author is Mika Waltari), the fictional character Sinuhe (who is a doctor), and a particular cover for Sinuhe the Egyptian, which has the keyword hieroglyphs.

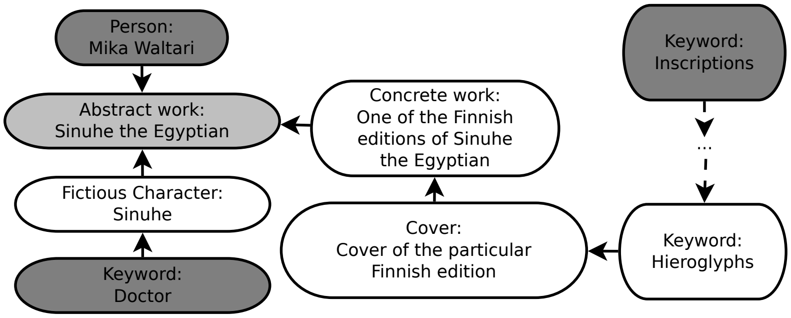


Fig. 3. Mapping to a final search result after text matching in BookSampo. Dark grey resources are those returned from label text matching, while the light grey resource is the final search result.

Finally, all resources that are not already abstract works must be mapped to any that they refer to, and finally an intersection taken between all abstract

[14] http://drupal.org/

works found to reveal the final result. Here, the character Sinuhe relates directly to the abstract work, but the cover is still two steps away. One must follow the link from the cover to the particular physical edition that it is the cover for, and from there finally to the abstract work.

To resolve this, the search functionality of CultureSampo was split into multiple stages, each taking in SPARQL queries. First, multiple "select" queries are run, one for each incoming keyword, acting on a dedicated CultureSampo index. Using this index, it is easy to efficiently return any resource that is related in any way to a literal or another resource with a label matching a particular text query. In addition, this index also performs subclass inference. Thus, from this stage, in the case of the example queries one would already get Sinuhe the Egyptian the book, Sinuhe and the cover, in addition to the more direct resource hits.

Then, a "mapping" query is run separately for each select query result set. In BookSampo, these map for example any returned covers, reviews, physical works, fictional characters and so on to the abstract works to which they refer. After this, the system automatically takes an intersection of the mapped results returned from each select query. A further "filter" query is also run. In BookSampo, this makes sure that only abstract books and authors ever make it to the final result set returned.

After the result set is finally obtained, it is paged and returned. This can still be manipulated by a "grouping" query. This can be used to ensure that for example a set amount of both authors and books matching a particular query are returned. To make sure all information to be shown in the search listing for each matched resource is included (such as cover images, years of first publication, etc.), the system still runs any given "describe" queries for each returned resource, before finally returning answers.

Because of the efficient indexes of CultureSampo as well as caching of e.g. the mapping query results, the average processing time for even these complex queries is still 100-400 milliseconds on a modern desktop server.

5 Discussion

Libraries have centuries of history in describing books as physical objects, particularly as pertains their physical location in the library. This leads to a large amount of institutional friction in applying new forms of indexing. For example, while libraries have talked of functional indexing (FRBR) from the early 1990s, actual systems have started to appear only concurrently with BookSampo.

Yet, before publishing the end-user portal, the benefits of using semantic web technologies in BookSampo have remained in part elusive to the library professionals. Particularly, there has been a noted scepticism with regard to the added value of ontologies versus the added cost of their maintenance. However, after the end-user portal was published, the search and recommendation functionalities afforded by the CultureSampo engine and the network model of information have been lauded as revolutionary, fulfilling the ideal model of fiction. For example, for a query of "Crime and Punishment", the system not only returns a single work, but actually places it in its literary historical context, also listing all authors that say they have been influenced or touched by the work, all other works that are compared to Crime and Punishment in their reviews, all kindred works and so on. Similarly, each work on its own page is automatically linked to other relevant works and the work's context by recommendation algorithms.

As far as the books and authors in BookSampo are concerned, they are also automatically integrated into the CultureSampo system with some 550,000 cultural objects in it. This makes it possible for the user of CultureSampo to approach the entire Finnish culture from a thematic starting point instead of starting with data type or a data producing organisation. For example, on can retrieve instantly data of museum objects, photographs, paintings, contemporary newspaper articles as well as literature dealing with, for example, agriculture in Finland in the nineteenth century. This way it is also possible, for example, to demonstrate the influences between different arts.

The present wish is to combine the BookSampo database semi-automatically with the Elonet database[15] of the Finnish National Audiovisual Archive, which would offer enormous amounts of additional data to both those in BookSampo who are interested in the film versions of books and those in Elonet who would like to know more about the source work of their favourite film.

Since the contents of BookSampo adhere to the principles of linked open data, they also automatically combine in a wider context with all other such material. For example, further information on both authors and books could be sourced from DBPedia. This way, BookSampo gradually approaches the entire context space of literature described in the ideal model for fiction, where "linking carries on ad infinitum".

There has also already been an example where the linked data of BookSampo could be used in a context outside the original environment it was designed for. On 23 May 2011, the major Finnish newspaper Helsingin Sanomat organized an open data hacking event, which utilized the BookSampo database through an inferring SPARQL endpoint. The analyses and visualization of the materials revealed, for example, that international detective stories have become longer since the beginning of the 1980s—from an average of 200 pages to 370 pages—but Finnish detective stories did not become longer until the 2000s. Other results combined BookSampo data with external grant data, showing for example what types of topics most likely receive grant funding or awards. Even new interactive applications were created, allowing users to discover which years were statistically similar from a publishing viewpoint, or locating all the places associated with Finnish fiction on a map.

Acknowledgements Thanks to Erkki Lounasvuori, Matti Sarmela, Jussi Kurki, Joeli Takala, Joonas Laitio, and many others for co-operation. This research is part of the National Finnish Ontology Project (FinnONTO) 2003–2012, funded by the National Technology Agency (Tekes) and a consortium of 38 companies

[15] http://www.elonet.fi/

and public organizations. The BookSampo project itself is funded by the Finnish Ministry of Education and Culture.

References

1. Davis, I., Galbraith, D.: Bio: A vocabulary for biographical information, http://vocab.org/bio/0.1/.html
2. Doerr, M.: The CIDOC CRM – an ontological approach to semantic interoperability of metadata. AI Magazine 24(3), 75–92 (2003)
3. Hickey, T.B., O'Neill, E.T., Toves, J.: Experiments with the IFLA functional requirements for bibliographic records (FRBR). D-Lib Magazine 8(9) (September 2002)
4. Hyvönen, E.: Developing and using a national cross-domain semantic web infrastructure. In: Sheu, P., Yu, H., Ramamoorthy, C.V., Joshi, A.K., Zadeh, L.A. (eds.) Semantic Computing. IEEE Wiley - IEEE Press (May 2010)
5. 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., Nyberg, K.: 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)
6. Hyvönen, E., Mäkelä, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., Junnila, M., Kettula, S.: MuseumFinland—Finnish museums on the semantic web. Web Semantics: Science, Services and Agents on the World Wide Web 3(2–3), 224–241 (Oct 2005)
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 (June 2010)
8. Mäkelä, E., Ruotsalo, T., Hyvönen, E.: How to deal with massively heterogeneous cultural heritage data – lessons learned in culturesampo. Semantic Web – Interoperability, Usability, Applicability (2011), accepted for publication.
9. Nelson, J., Cleary, A.: FRBRizing an e-library : Migrating from dublin core to FRBR and MODS. code{4}lib (12) (December 2010)
10. Riva, P., Doerr, M., Zumer, M.: FRBRoo: enabling a common view of information from memory institutions. In: World Library and Information Congress: 74th IFLA General Confrence and Council (Aug 2008)
11. Saarti, J.: Aspects of Fictional Literature Content Description: Consistency of the Abstracts and Subject Indexing of Novels by Public Library Professionals and Client (in Finnish). Ph.D. thesis, University of Oulu (November 1999)
12. Saarti, J., Hypen, K.: From thesaurus to ontology: the development of the kaunokki Finnish fiction thesaurus. The Indexer 28, 50–58(9) (June 2010)
13. Serola, S., Vakkari, P.: Yleinen kirjasto kuntalaisten toimissa; Tutkimus kirjastojen hyödyistä kuntalaisten arkielämässä. Finnish Ministry of Education and Culture (May 2011)
14. Tran, T., Cimiano, P., Rudolph, S., Studer, R.: Ontology-based interpretation of keywords for semantic search. In: The Semantic Web. pp. 523–536. Springer (2007)
15. Viljanen, K., Tuominen, J., Hyvönen, E.: 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, Proceedings. pp. 781–795. Springer-Verlag (2009)