Improving Fiction Literature Access by Linked Open Data -Based Collaborative Knowledge Storage - the BookSampo Project

Eetu Mäkelä (1)
Kaisa Hypén (2) and
Eero Hyvönen (1)
[1]: Aalto University, [2]: Turku City Library
Finland

Session:

141 — *Potential of knowledge management in public libraries* — Knowledge Management

Abstract:

BookSampo is a joint project between the Finnish public libraries and semantic web researchers, to improve fiction literature search and recommendation. In the project, dozens of librarians around Finland have used a collaborative web-based metadata editor to input diverse knowledge about fiction literature into a shared database. Particularly, the project has sought to improve access by indexing not only bibliographical information about the books, but focusing on the content and context of the works. In order to do this, the database employs advanced techniques such as functional, content-centered indexing, ontological vocabularies and the networked data model of linked open data. To demonstrate the functionality this makes possible, the fiction literature portal http://www.kirjasampo.fi/ was created. This portal uses the knowledge created in the project to offer advanced semantic search and recommendation based on the database created. In addition, web services exposing direct access to the data have been used for example in culture hack events to answer more complex questions, such as where in Finland are the most crimes committed in fiction literature.

1 Introduction

With the advent of the Internet, the role of public libraries as primary sources of factual knowledge has diminished, particularly among younger people. This is reflected in the analyses published in a 2011 study of public library use in Finland [18]. 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, libraries are still an important source for fiction literature. While 34% of the respondents said they still benefited from factual information stored in libraries, the percentage for fiction was 45%.

These results encourage libraries to improve their services related to fiction literature, such as search and recommendation facilities. However, the nature of fiction necessitates a move from 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 entered only since 1997, using the fiction content thesaurus Kaunokki[1], developed since 1993.

On the other hand, already in 1999 a study [16] on fiction literature indexing concluded that customers' descriptions of the pertinent points of, and questions about fiction 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 publication data on the book and content description, but also 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 2008 started a venture to develop new ways of describing fiction, the model was chosen as a concrete goal. Because the model placed emphasis on the connections between information items, it seemed a good fit for semantic web technologies. Thus, the libraries approached the Semantic Computing Research Group at Aalto University and the University of Helsinki, who had prior experience in publishing cultural heritage content on the semantic web, having created the MuseumFinland portal [8] in 2004 and the CultureSampo portal [7 ,13] in 2009. Soon, the BookSampo project started as part of the national FinnONTO initiative[2].

Today, the end-user portal created in the project at http://www.kirjasampo.fi/ provides access to virtually all fiction literature published in Finland since the mid 19th century, some 90 000 works, 30 000 authors and 2 500 publishers.

In the following, lessons learned during the BookSampo project will be presented. First discussed are the many insights gained in modelling fiction. Then the paper presents the parts of the system, focusing on the challenges faced and benefits gained from applying semantic web technologies. Finally, the paper discusses the reception of the developed system in library circles.

1 http://kaunokki.kirjastot.fi/

2 http://www.seco.tkk.fi/projects/finnonto/

2 The BookSampo Data Model

From the start, the BookSampo project was geared towards an ambitious, disruptive experiment as opposed to an incremental improvement. Thus, it was decided that the MARC format[3] 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 current library indexing systems. The reason for this was the recognition that the MARC format, and the systems built on it exhibit a core problem regarding indexing the rich cultural context into which the works belong.

This core problem stems from the tradition of cultural heritage institutions to index content in schemas where only a single primary content type is modeled as an object, with everything else referred to and described only as text strings. As an example, library databases typically contain only books as objects, while publishers and authors are entered as text. On the other hand, in the authority database of the same library, the actors such as authors and publishers are the objects, with their details as text. Because of this, anyone wanting to combine these datasets must themselves utilize imperfect text matching techniques.

In essence this approach dictates a series of flat, narrow, pre-selected single points of view into the world, counterproductive to revealing the context to which the works belong. For example, for situating a work within its wider context, any information available on its author may be important. Do authors with different backgrounds write on different themes? Do authors from the same locale form (thematic) cliques? What do authors who received literary awards or government grants write about? In order to support queries such as these, a vast amount of information needs to be stored pertaining to the authors, while also allowing for error-free matching of this author information to the information on the books.

It makes no sense to copy all this information on the authors into every book records in the book database, but it is also too resource intensive and error prone to match between different databases using string comparison techniques.

Thus, the database chosen for the BookSampo project had to itself be able to contain all the different types of objects, as well as relate them to each other unambiguously. For this, the RDF data model [10], is an exceptional fit.

3 http://www.loc.gov/marc/

2.1 Introduction to the RDF data model

RDF Source A

Subject	Predicate	Object
http://bs.fi/Sinuhe	http://bs.fi/author	http://bs.fi/Waltari
http://bs.fi/Sinuhe	http://bs.fi/character	http://bs.fi/Pharaohs
http://bs.fi/Sinuhe	http://bs.fi/character	http://bs.fi/SinuheC
http://bs.fi/Sinuhe	http://bs.fi/name	“Sinuhe the Egyptian”@en
http://bs.fi/Sinuhe	http://bs.fi/name	“Sinuhe egyptiläinen”@fi
http://bs.fi/SinuheC	http://bs.fi/name	“Sinuhe”
http://bs.fi/Waltari	http://bs.fi/name	“Mika Waltari”

RDF Source B

Subject	Predicate	Object
http://bs.fi/Waltari	http://bs.fi/member	http://bs.fi/Tulenkantajat
http://bs.fi/Tulenkantajat	http://bs.fi/description	“A Finnish literature group”

RDF Source C

Subject	Predicate	Object
http://bs.fi/SinuheC	http://bs.fi/occupation	http://bs.fi/Doctor

Table 1**:** Example triples

The RDF data model is based on storing all data as triples of the form <subject,predicate,object>, with each slot in a triple being a reference to a globally unique URI identifier, or alternatively in the case of the object position, to a descriptive, possibly language-coded literal.

For example, consider the triples found in RDF Source A in Table 1. These triples tell that the author of (the book) Sinuhe the Egyptian is Mika Waltari, and that it has a pharaoh character, as well as a character named “Sinuhe”. There are no inbuilt restrictions in the RDF data model on what can URIs can be put in each position, nor hardly any meaning assigned to particular URIs. However, because each URI is a global identifier for the thing it represents, and any URI can be put into the subject position, it is always possible to give more information on any object. This results in a directed graph, as visualized for the example triples in Figure 1. It should also be noted how the use of URI identifiers supports language neutrality, as labels for any object can be added in any language desired, such as in the case of the book “Sinuhe the Egyptian”/”Sinuhe egyptiläinen” in the example.

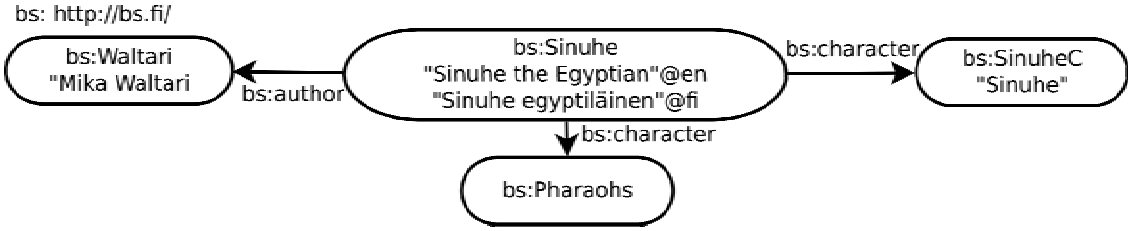

Figure 1: Example triples from source A as a directed graph

Now, suppose that also the RDF Sources B and C of Table 1 are available. Because the URI identifiers used in RDF are *global*, any RDF utility loading these now sees the data graph as depicted in Figure 2. This instant data integration is at the core of the RDF data model.

Figure 2: Combined example triples as a directed graph

With RDF being based on every entity being a globally referenceable object, and operating on an a priori unbounded set of properties and relations between them, the RDF data model seemed an exceptional match for realizing the complex network of cultural context into which books and authors belong.

2.2 Introduction to ontologies and inference

Aside from the RDF data model, the second major cornerstone of the semantic web technology stack are ontologies and the inferencing capabilities they offer. In the semantic web context, an ontology can be defined as a formal, explicit specification of a shared conceptualization [4], usually in the context of a particular domain of knowledge. In practice, they are basically just files containing more RDF triples. Their power comes from the fact that the properties used in them are defined in ontology languages such as RDFS [1], OWL [11] and SKOS [12], which also describe what can be inferred based on these properties.

For example, the following rules are commonly agreed-upon:

- X rdf:type Y $\rightarrow$ X is an individual of class Y. Everything that holds true for Y holds true for X

- X rdfs:subClassOf Y → X is a subclass of Y, Everything that holds true for Y holds true for X
- X owl:sameAs Y → X and Y are the same thing. Everything that holds true for Y holds true for X and vice-versa.

Now, taking these rules, consider what can be inferred by adding the extremely simple ontology and mappings listed in table 2 to the triples discussed before. For example, now a search making use of inferencing will match the book "Sinuhe the Egyptian" not only when searching for books with pharaohs as characters, but any rulers in general, and also match the book through the character Sinuhe to any queries for books with doctors or medical workers. A recommendation system on the other hand could now recommend as similar not only other books about pharaohs, but also books about other rulers such as dictators.

In addition to supporting inferencing, standardized general ontologies containing common concepts are also often used as a semantic glue to bind together multiple RDF datasets. For example, supposing that the example ontology is either directly used in other datasets or can be mapped to their ontologies, it becomes instantly possible for a semantic web recommendation system to recommend not only other books with ruler characters, but for example films, pictures, poems or any other object from another compatible RDF dataset.

Ontology

Subject	Predicate	Object
onto:Doctors	rdfs:subClassOf	onto:Medical_Workers
onto:Pharaohs	rdfs:subClassOf	onto:Rulers
onto:Dictators	rdfs:subClassOf	onto:Rulers

Mappings

Subject	Predicate	Object
bs:occupation	owl:sameAs	rdf:type
bs:Pharaohs	owl:sameAs	onto:Pharaohs
bs:Doctor	owl:sameAs	onto:Doctors

Table 2**:** Example ontology and mappings

2.3 Applying Ontologies in BookSampo

As stated, ontologies can be used to increase the intelligence of semantic web search and recommendation systems. So, to maximally leverage the possibilities of semantic web technologies in the project, an ontology infrastructure was needed that could cater to the content and context of fiction literature, and hopefully also act as a glue tying the data to its wider cultural heritage context.

Here, the project leveraged the work done in the wider FinnONTO project [6], 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 combined ontology 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[4]. 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. Mostly these are lightweight ontologies converted from prior existing thesauri. This has the benefit that there is already a long tradition of indexing content with these thesauri, which can then immediately be made use of, instead of having to reindex old content.

The KOKO ontology cluster however lacked many concepts relevant to indexing the content of fiction literature, such as themes, genres and milieus. Luckily, proven paths could be followed here because of the existance and use of Kaunokki and Bella, the Finnish and Swedish thesauri for fiction indexing. In the BookSampo project, these were converted into the bilingual ontology KAUNO [17].

What is actually done in converting thesauri into light-weight ontologies in the KOKO infrastructure is to examine, correct and extend the broader and narrower term links in the source thesaurus, so that a proper subsumption hierarchy of concepts is formed, suitable for machine inferencing. The experience of the librarians who ontologized Kaunokki was that this brought in a very welcome additional structuring to the vocabulary.

For example, the term “american dream”, in the Kaunokki 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.

As for linking the KAUNO ontology to KOKO, thus far, with the exception of MAO and TAO, the KOKO ontologies are each joined only through common links to the national upper ontology 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[5]. All automatically created links were checked by hand, as well as new links created.

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. Also, by linking this new ontology to KOKO, all material indexed using Kaunokki

4 http://www.yso.fi/onki3/en/overview/yso

5 http://protege.stanford.edu/

and Bella were immediately bridged to all the other cultural heritage already indexed using other KOKO constituents, for example in the CultureSampo portal.

In addition to KOKO, the project also makes use of other resources for picking rich and interlinked indexing terms. These are 1) the LEXVO language ontology[6], 2) the Getty Union List of Artist Names[7] with different spellings of artists' names, birth and death information, contact information and so on, and 3) a unified place ontology termed KOKO-Place, which includes 17 million locations with coordinate information gathered from sources such as GeoNames[8], OpenStreetMap[9] and the National Land Survey of Finland place name database.

2.4 Data Modeling in BookSampo

While the RDF data model itself allows one to say practically anything about anything, for a particular database and application it still makes sense to stick to a relatively uniform schema. This schema can however grow and change at will, a schema description being actually just more RDF triples with particular property URIs that have been globally agreed to describe schemas. Such schema shift also happened in the case of the BookSampo project.

In BookSampo, the objects with the most properties defined in the schema currently are books and authors. For authors, the schema currently defines nineteen reference properties and five literal properties, listed in Table 3. As can be seen, a lion's share of the properties are object references. First, this allows additional information to be given for each reference, such as coordinates for the locations, superclasses for the keywords, display labels for the concepts in different languages and so on. Second, it allows reusing these objects in describing other authors or books, instantly making this additional information also available in relation to them.

6 http://www.lexvo.org/

7 http://www.getty.edu/research/tools/vocabularies/ulan/index.html

8 http://www.geonames.org/

9 http://www.openstreetmap.org/

Reference Property	Source
Occupation	KOKO ontology, 126 in-project additions
Gender	Two in-project resources
Mother tongue	LEXVO language ontology
Nationality	Getty ULAN nationalities, 67 in-project additions
Is same person as	Other actors in the project (Allows keeping pen-names separate, yet keeps the identities linked)
Author's picture	Picture description resources in the project
Time of birth	Date resources in the project
Place of birth	KOKO-Place ontology, 594 in-project additions
Place of education	KOKO-Place ontology, 594 in-project additions
Place of residence	KOKO-Place ontology, 594 in-project additions
Time of death	Date resources in the project
Place of death	KOKO-Place ontology, 594 in-project additions
Education	KOKO ontology
Has award	Award resources in the project
Associated schools, style periods	30 in-project resources
Positions of trust, memberships	124 in-project resources
Hobbies	KOKO ontology, 18 in-project additions
Source and reference links	Link description resources in the project
Regional library area	Regional library areas in the project
Cataloguer	Actor resources in the project

Literal Property
Name
Alternative name
Biographical text
Writer's own words
Additional information
Text sample

Table 3**:** Properties for authors stored in the database

As also evident, while preferring the use of shared ontologies from which to draw concepts from, the project also allows indexers to add new concepts when these shared repositories fall short. However, even these local terms need not lie alone in the shadows, as they can be linked to the existing ontology framework through defining their ontological superclass. This way for example, the occupation "specialized nurse" which was lacking in the common KOKO hierarchy could be added, while still linking it to the "nurses" concept in the ontology, and through that also to the other medical staff already there.

As stated, 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 as events, in the spirit of the cultural heritage interchange model of CIDOC-CRM [3] and the BIO-schema of biographical information [2].

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.

This model of a single object worked well, until it was decided that the project should also extend to include Swedish literature[10], 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 [15], 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 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

10 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.

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. For a listing of the fields used to describe an abstract work, see Table 4, while the fields relating to the physical work are in Table 5. In the end-user interface, the data on any and all physical works linked to an abstract work are shown 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.

Reference Property	Source
Creator	Actor resources in the project
Type	literary type resources in the project
Genre	KAUNO ontology genre facet
Theme	KAUNO ontology theme facet
Character types in the narrative	KAUNO ontology character type facet
Main character	Character resources in the project
Place of events keyword	KAUNO ontology place type facet
Concrete place of events	KOKO-Place, 594 in-project additions
General time of events	KAUNO ontology era facet
Concrete time of events	Date resources in the project
Keyword	KAUNO ontology, 1034 in-project additions
Combined keyword	Combined keyword resources in the project
Physical works	Physical work versions of the book in the project **or** parts of physical works
Original language	Languages in the LEXVO ontology
Has award	Award resources in the project
Films and other adaptations	Other work resources in the project
Librarian recommends	Other related work resources in the project
Fulltext links	Link description resources in the project
Source and reference links	Link description resources in the project
Reviews	Review description resources in the project
Cataloguer	Actor resources in the project

Literal Property
Name
Alternative title
Textual description
Text sample

Table 4**:** Properties for abstract works stored in the database

Reference Property	Source
First publication	Boolean resource in the project
Original work	Link to the original physical work if a translation
Language	LEXVO language ontology
Publisher	Publisher resources in the project
Year of publication	Date resources in the project
Cover	Cover description resources in the project
Translator	Actor resources in the project
Illustrator	Actor resources in the project
Other creator	Actor resources in the project
Part of series	Part of series resources in the project
Part of physical work	Physical work resources in the project

Literal Property
Name
Subtitle
Number of pages
Complementary information about publication history

Table 5**:** Properties for concrete works stored in the database

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[11] recognizes work and edition levels, with the latter also combining expression and manifestation. Exactly the same situation is present also in the LibraryThing portal[12], 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[13], and the OCLC WorldCat system[14], 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

11 http://www.openlibrary.org/

12 http://www.librarything.com/

13 http://www.australianmusiccentre.com.au/about/websitedevelopment

14 http://frbr.oclc.org/pages/

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 [5 ,14], so should a need for further separation arise, one could just repeat a similar split procedure 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 3.

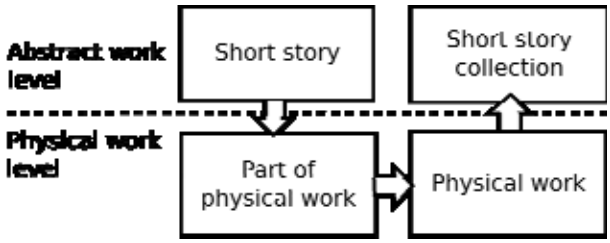

Figure 3: Relationship between short story and short story collection in the BookSampo data model.

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

As can be seen from the tables describing the fields of the abstract and physical works, both their content and context are richly described in BookSampo. However, to fully appreciate the network formed, one must also look at what information is given about the secondary resources mentioned. The details about the author, of course linked to all their works, have already been discussed. The schemas of the other secondary resources are described in Table 6.

Picture

Literal properties:	Name, URL, Description

Book Cover

Literal properties:	Name, URL, Description
Reference properties:	Illustrator (Actor resources in the project), Keyword (KAUNO ontology, combined keywords in the project, local keywords)

Keywords Belonging Together (e.g. suicide : justification)

Literal properties:	Name
Reference properties:	Keyword (KAUNO ontology, in-project additions)

Web Link

Literal properties:	Name, Description, URL

Fictional Character

Literal properties:	Name, Description
Reference properties:	Keyword (KAUNO ontology, in-project additions), Personification of (Actors in the project. This way, real persons and their versions in fiction can be kept separate, yet linked)

Award

Literal properties:	Name
Reference properties:	Award Series (Award Series in the project), Award Year (Dates in the project)

Award Series

Literal properties:	Name, Alternate name, Description
Reference properties:	Keyword (KAUNO ontology, in-project additions)

Part of Series

Literal properties:	Name, Number in Series
Reference properties:	Series (Series in the project)

Series

Literal properties:	Name, Description
Reference properties:	Keyword (KAUNO ontology, in-project additions)

Literary School or Period

Literal properties:	Name, Description
Reference properties:	Concrete timespan (Dates in the project), Concrete place (KOKO-Place ontology, in-project additions), Keyword (KAUNO ontology, in-project additions)

Position of Trust

Literal properties:	Name, Alternate name, Description
Reference properties:	Keyword (KAUNO ontology, in-project additions)

Place	
Literal properties:	Name, Description, Latitude, Longitude
Reference properties:	Larger Place (KOKO-Place ontology, in-project additions)
Date	
Literal properties:	Name, Earliest Possible Start, Latest Possible Start, Earliest Possible End, Latest Possible End (ISO 8601 dates)

Table 6**:** Properties for secondary resources

Here is evident also 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 auxiliary resource where this information can be recorded. In the BookSampo schema, for example, to say that a book is part 7 in the yellow library series, one must relate it to the "part of series" resource "Part 7 in the Yellow Library", which is in turn annotated as a part of the "Yellow Library series" resource and having 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 our system, 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.

On the other hand, as a testimony 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.

3 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 end-user portal search and browsing functionality is discussed.

3.1 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[15] 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.

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

15 http://www.helmet.fi/search□S9/

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.

3.2 Collaborative Semantic Web Editing Environment

As BookSampo was to base its database natively on RDF, the project decided to adopt the SAHA RDF-based metadata editor [9] developed by the FinnONTO project as its primary editing environment.

SAHA is a general-purpose, adaptable editing environment for RDF data. It centers on projects, which contain the RDF data of a single endeavour. The main screen of SAHA provides a listing of the object types 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 4. 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 [19], 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 4 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.

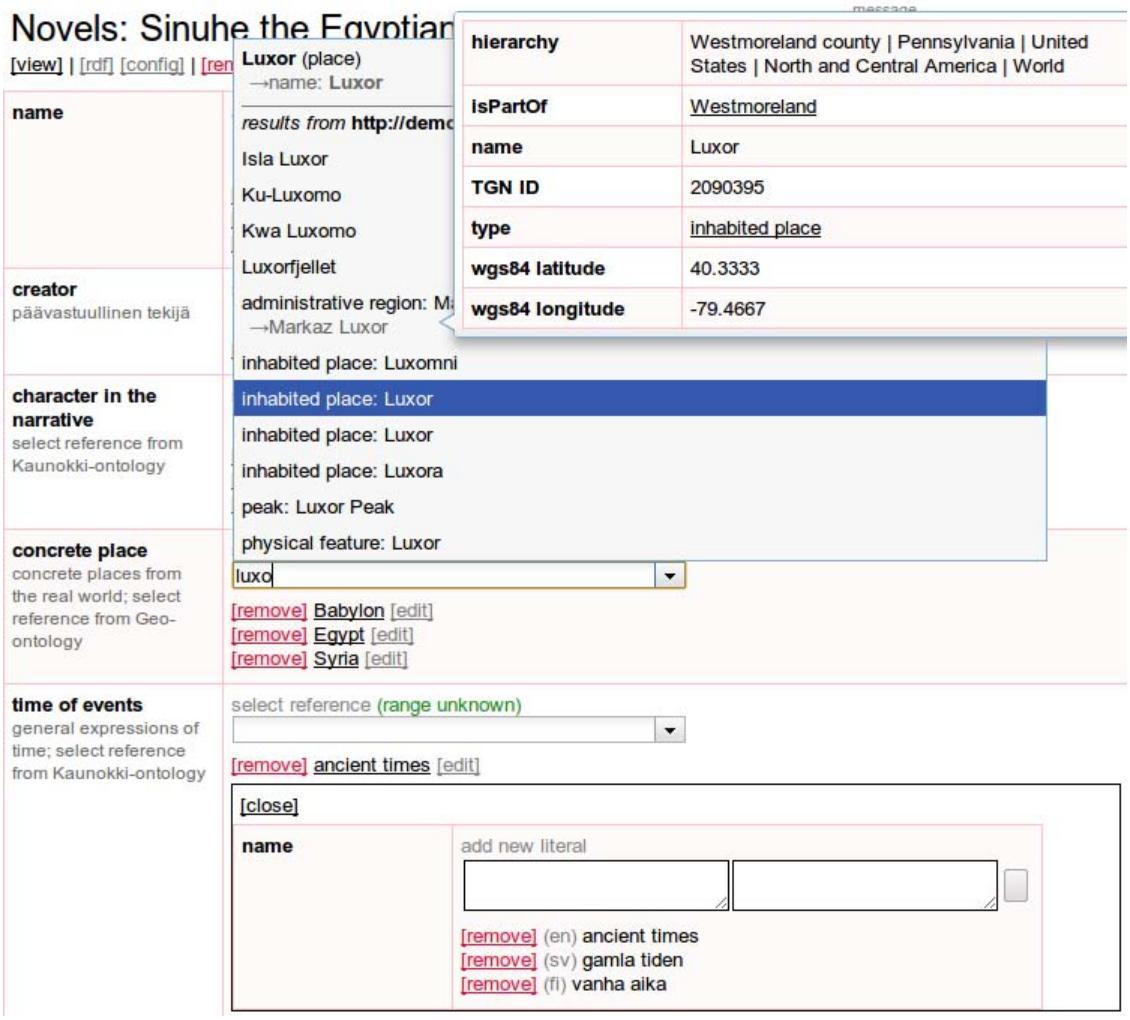

Figure 4: The SAHA metadata editor, showing both semantic autocompletion as well as a pop-up preview presentation of one of the autocompletion results.

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 view. 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 4, 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 one 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 4 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 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 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 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 indexers taking part in BookSampo indexing have found the SAHA editor to be intuitive, 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.

3.3 End-User Portal

The end-user interface of BookSampo, available at http://www.kirjasampo.fi/ is a separate application, built on top of the Drupal[16] portal system. However, all primary information is kept and served by a semantic web -enabled back-end system, with the Drupal layer only adding its ready-made commenting and tagging functionality, forums, blogs etc. on the client side.

The two main basic functionalities provided by the front-end portal are searching and browsing of the repository. In the search interface, text queries are passed along to the back-end along with patterns determining how content is matched. For example, the plain text query “Waltari Doctor Inscriptions” would first locate Mika Waltari the actor and the Doctor and Inscriptions KOKO ontology concepts. Then, utilizing various mapping patterns (e.g. any character types matching the search are mapped through characters of those types back to the books the character appears in), the abstract work “Sinuhe the Egyptian” can be returned. Such complex matches can and need also be explained to the user, so that they know e.g. that the work matches, because Waltari is the book’s author, because it has a main character who is a doctor

16 http://drupal.org/

and because one of its editions has a cover that contains hieroglyphs, which are a type of inscription.

The browsing interface on the other hand provides the user a possibility to wander through the context of a work and through it to other works. Besides simply allowing one to walk the semantic network through the actors, books and keywords, the interface also provides semantic recommendations, which automatically locate interesting semantically related content in some way related to the currently viewed work. For example, for the Sinuhe book, the system recommends Nefritite by André Chedid, with an explanation that both are historical novels dealing with the way of life in Egypt in the 13th century BC.

By combining the Drupal repositories of ready-made modules with the rich query functionality the back-end provides, it has also been possible to iteratively and quickly add new functionalities to the front-end as needed. Such are for example creating a tag cloud of random content terms for serendipidous querying, creating a functionality where users can gather literary works to virtual shelves, which they can then share with each other, as well as providing custom views to for example the book covers and contemporary reviews indexed in the system.

4 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 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.

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, the semantic web version of Wikipedia. This way, BookSampo gradually approaches the entire context space of literature described in the ideal model for fiction, where "linking carries on ad infinitum".

The linked data of BookSampo has also already been 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 a web service 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. 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 and public organizations. The BookSampo project itself is funded by the Finnish Ministry of Education and Culture.

References

[1] Brickley, D., Guha, R.V.: Resource description framework (RDF) schema specification (2000), w3C Candidate Recommendation http://www.w3.org/TR/rdf-schema

[2] Davis, I., Galbraith, D.: Bio: A vocabulary for biographical information, http://vocab.org/bio/0.1/.html

[3] Doerr, M.: The CIDOC CRM – an ontological approach to semantic interoperability of metadata. AI Magazine 24(3), 75–92 (2003)

[4] Gruber, T.R.: A translation approach to portable ontology spesifications. Knowledge Acquisition 5(2), 199–220 (1993)

[5] 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)

[6] 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)

[7] 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)

[8] 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)

[9] 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)
[10] Manola, F., Miller, E. (eds.): RDF Primer. The World Wide Web Consortium (February 2004), http://www.w3.org/TR/rdf-primer/, w3C Recommendation
[11] McGuinness, D.L., van Harmelen, F. (eds.): OWL Web Ontology Language Overview. World Wide Web Consortium (Feb 2004), http://www.w3.org/TR/owl-features/, w3C Recommendation
[12] Miles, A., Bechhofer, S. (eds.): SKOS Simple Knowledge Organization System Reference. World Wide Web Consortium (Aug 2009), http://www.w3.org/TR/skos-reference/, w3C Recommendation
[13] 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.
[14] Nelson, J., Cleary, A.: FRBRizing an e-library : Migrating from dublin core to FRBR and MODS. code{4}lib (12) (December 2010)
[15] 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)
[16] 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)
[17] Saarti, J., Hypen, K.: From thesaurus to ontology: the development of the kaunokki Finnish fiction thesaurus. The Indexer 28, 50–58(9) (June 2010)
[18] Serola, S., Vakkari, P.: Yleinen kirjasto kuntalaisten toimissa; Tutkimus kirjastojen hyödyistä kuntalaisten arkielämässä. Finnish Ministry of Education and Culture (May 2011)
[19] 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)