Undefined 0 (0) 1
IOS Press

Fiction Literature as Linked Open Data —the BookSampo Dataset

Eetu Mäkelä [a,*], Kaisa Hypén [b], and Eero Hyvönen [a]

[a] *Semantic Computing Research Group (SeCo), Aalto University and University of Helsinki, Espoo, Finland*
[b] *Turku City Library, Turku, Finland*

Abstract.
The BookSampo dataset provides information as linked data on fiction literature published in Finland going back to the 15th century, along with rich descriptions of both their content and context. The dataset contains data on nearly 400,000 subjects, including literary works, authors, book covers, reviews, awards, images, and movies, over 3 million triples in total. The data has been applied as the basis of the BookSampo portal in public use in Finland, and is aligned with the cross-domain cultural heritage contents and ontologies of CultureSampo, another in-use semantic portal. The data has been used to answer complex questions, such as what topics should one write about, if one wants to get a literary award (based on statistics). The metadata was transformed into RDF from legacy library databases, then enriched manually by dozens of librarians in a Web 2.0 fashion in Finnish public libraries, and is constantly updated at a rate of some new 90,000 triples monthly.

Keywords: linked data, fiction literature, dataset, modelling

1. Introduction

The role of public libraries as a source for fiction literature remains strong even if their role as a provider of factual knowledge has diminished [11]. This encourages libraries to improve their services related to fiction literature. However, the nature of fiction necessitates a departure from old library indexing traditions. Classifying books e.g. by their genre and shelf location alone is not enough: studies on user needs in fiction literature show that satisfactory fiction literature indexing systems must also model and store the rich connections between fiction literature works, their authors, and their surrounding cultural context [10].

The BookSampo project is a joint venture by the Finnish public libraries and semantic web researchers to provide such a system using Linked Data. For the end-users of public libraries, the project manifests as http://www.kirjasampo.fi/, a portal for finding and browsing rich interlinked information about fiction [6]. This paper however describes the dataset underlying the system, an experiment in developing and using rich Linked Data as a basis for fiction literature services.

2. Core Dataset Information

The BookSampo dataset is located on the web in an instance of the SAHA RDF-based metadata editor [5]. SAHA is used by the librarians for maintaining the dataset, and it also provides a SPARQL end point to the data, as well as RDF export functionality. The relevant web addresses are:

SAHA Browser & Editor View
http://saha.kirjastot.fi/kirjasampo/index.shtml
RDF Export
http://saha.kirjastot.fi/kirjasampo/export.shtml
SPARQL Endpoint
http://saha.kirjastot.fi/service/data/kirjasampo/sparql

The scope of the BookSampo data is all fiction literature published in Finland in Finnish or Swedish[1], be they original or translated. Originally, the data in

*Corresponding author. E-mail: eetu.makela@aalto.fi

[1] Finland is a bilingual country.

BookSampo came from a June 2009 dump from the Helsinki metropolitan area library system, converted into RDF. Newly published material after that is downloaded and automatically converted each night from BTJ Finland Ltd, a company that provides these descriptions to Finnish library systems for a price. However, both of these data sources are still very bibliographically oriented. The value of BookSampo comes from the fact that dozens of librarians around Finland constantly collaboratively edit and improve the data loaded into BookSampo using the SAHA metadata editor. This effort has made the sparse metadata about older literature in the library systems richer, and more importantly, extended the metadata to cover related objects, such as awards, publishers, reviews etc. For this purpose, additional datasets (authority databases, book covers, reviews, etc.) were also converted into RDF and aligned with the basic bibliographic records.

Unfortunately, the diverse origins of the data in BookSampo create problems in unambiguously licensing the complete dataset. The data of the Helsinki metropolitan area library system, still making up the bulk of the combined dataset, has been released under the Creative Commons Attribution-ShareAlike 1.0 licence. For the other library databases transformed and the manual work done by the librarians, clear licencing was never agreed upon, other than an informal agreement of no strings attached. However, the company BTJ Finland Ltd does claim a right on the annotations flowing in from their system. Thus, when the dataset has previously been used by outsiders, either all data on literary works since June 2009 has been evicted, or permission for the use has been sought from BTJ.

At the time of writing, the dataset contains a little over 3 million triples, pertaining to nearly 400,000 subject URIs. Nearly 2 million of the triples have URIs as property values (objects), testifying to the richness of the semantic network formed. The dataset currently grows with a rate of approximately 90,000 triples and 10,000 subject URIs per month. Of the literals in the RDF database, some 560,000 are without a language code, consisting of, e.g. person names, dates and URLs. Some 260,000 literals are in Finnish, 130,000 in Swedish, and 36,000 in English, with other languages evident in much smaller numbers.

3. Contents of and Links Outside the Dataset

The important classes in use in BookSampo along with their approximate instance counts are given in Table 1. Besides these core classes, there are a number of auxiliary types for the ranges of annotation fields, such as professions, schools, and time periods. In addition to local resources, the project also makes use of external vocabularies for these fields. These also act as semantic glue that allows the project to reach past its own boundaries into other linked datasets, and into the wider cultural historical context it seeks.

Type	Number of Instances
Literary Works	93,000
Editions	127,000
Book Covers	27,000
Fictional Characters	19,000
Contemporary Reviews	15,000
Weblinks	10,000
Literary Series	2,900
Literary Awards	2,700
Literary Award Series	200
Movies	1,100
People (e.g. Authors)	29,000
Author's Pictures	2,600
Publishers	2,600

Table 1
Important classes in BookSampo along with their instance counts

Most external references are to the KOKO lightweight class ontology [4], a central part of the Finnish semantic web infrastructure, currently comprised of 14 domain ontologies joined under the Finnish national upper ontology YSO[2]. A particularly used part of KOKO is the KAUNO fiction content indexing ontology, which was transformed from the Finnish-Swedish thesaurus Kaunokki-Bella[3] into RDF, changed manually into a simple RDF(S) ontology, and aligned with YSO for use in the BookSampo project.

In addition to KOKO, the project also makes use of the LEXVO language ontology[4], the Getty Union List of Artist Names[5] with different spellings of artists' names, birth and death information, and so on, and a unified place ontology termed KOKO-Place, which includes 17 million locations with coordinate information gathered from sources such as GeoNames[6], DB-

[2] http://www.yso.fi/onki3/en/overview/yso
[3] http://www.kirjastot.fi/fi-FI/kaunokkibella/
[4] http://www.lexvo.org/
[5] http://www.getty.edu/research/tools/vocabularies/ulan/index.html
[6] http://www.geonames.org/

Pedia[7], OpenStreetMap[8], and the National Land Survey of Finland place name database[9].

Looking at the dataset from a viewpoint of linking to the dataset from an international context, promising points of entry are the places and the actors, as they are not language bound and already to an extent use URIs from international datasets. As for the KOKO class ontology, there are some, mostly automatically created mappings to, for example, the Getty AAT thesaurus, Wordnet, and DBPedia. However, their quality has not yet been evaluated. For creating such mappings by oneself, the YSO upper ontology (ca. 25,000 concepts) of KOKO has labels for virtually all concepts in Finnish, Swedish, and English, while the KAUNO part has labels primarily only in Finnish and Swedish.

The BookSampo dataset has grown organically over several years using different tools, so the user should be aware of a few conventions, or lacks thereof: First, practices and languages vary regarding minting URIs for classes, instances, and properties. However, all the schema properties and classes have labels in Finnish, Swedish, and English. Second, the schema definitions in the dataset virtually violate RDFS semantics in one major aspect, due to the specifics of the SAHA editor used: properties may have multiple separate domain and range constraint statements, but this doesn't imply that the instances related by these properties are members of the intersection of domain/range classes, as required in the RDF Schema specification. Instead, the domain/range definitions are used in SAHA with the interpretation that the union of the class restrictions applies. For example, dc:description may then have as its domain restriction two non-intersecting classes for literary works and authors.

4. The BookSampo Data Model

The central objects in BookSampo, around which the others cluster, are books and authors. For authors, the schema currently defines nineteen object properties and five literal properties, listed in Table 2. A lion's share of properties have object references as values.

Table 2 also shows how both shared ontologies as well as local additions are used for drawing concepts for property values. However, any local terms should be linked to the ontology framework through defining their ontological superclass. This way, for example, the occupation "specialized nurse" which was lacking in the shared KOKO ontology, could be added, while still linking it to the "nurse" concept in the ontology, and through that to the other medical staff already there.

The author schema has gone through one major revision during the project: the way of encoding biographical information was changed from events to attributes. Initially, details about, among others, the times and places of authors' births, deaths and studies were represented in BookSampo as events, in the spirit of the cultural heritage interchange model 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

Object Property	**Source of Value**
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 additions
Place of education	KOKO-Place ontology, 594 additions
Place of residence	KOKO-Place ontology, 594 additions
Time of death	Date resources in the project
Place of death	KOKO-Place ontology, 594 additions
Education	KOKO ontology
Has award	Award resources in the project
Associated schools, periods	30 in-project resources
Positions of trust, memberships	124 in-project resources
Hobbies	KOKO ontology, 18 in-project additions
Reference links	Link description resources in the project
Regional library	Regional library areas in the project
Cataloguer	Actor resources in the project
Literal Properties	
Name, Alternative names, Biographical text, Writer's own words, Additional information, Text sample	

Table 2

Properties for authors stored in the database

[7] http://dbpedia.org/

[8] http://www.openstreetmap.org/

[9] http://www.maanmittauslaitos.fi/aineistot-palvelut/rajapintapalvelut/nimiston-kyselypalvelu-wfs/kayttoonotto/aineistot-tuotteet

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.

The project thus changed back to a more traditional model, where data about times and places of occurrences are directly saved as author, not event attributes. In the case of representing degrees attained by authors, this did lead to some loss of data, since the flat attributes allowed only representation of multiple degrees without dates. However, the librarians deemed the simplicity to outweigh the costs in this situation.

In the BookSampo schema for literary works, it was important to be able to describe both properties relating to the content of the work, which stays the same in all translations and editions, and at the same time also to index edition-specific information, such as translators, publishers, and publication years. Here, the project draws from the FRBRoo Model [9], 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 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 flat 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 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: 1) as an abstract work, which refers to the contents of the work, which is the same in all translations and versions, and 2) as a physical work, which describes features inherent to each translation or version. 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. For a listing of the fields used to describe an abstract work, see Table 3, while the fields relating to the physical work are presented in Table 4.

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 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 library records to FRBR. For example, the Open library[10] recognizes work and edition levels, with the latter also combining expression and manifestation. Exactly the same situation is present also in the LibraryThing portal[11], 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[12], and the OCLC WorldCat system[13], 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.

It has also already been established by others that separation of expressions from even originally combined fields is possible by mostly automated process-

[10] http://www.openlibrary.org/
[11] http://www.librarything.com/
[12] http://www.australianmusiccentre.com.au/about/websitedevelopment
[13] http://frbr.oclc.org/pages/

Object Property	**Source of Values**
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	KAUNO ontology character type facet
Main character	Character resources in the project
Place of events	KAUNO ontology place type facet
Concrete place	KOKO-Place, 594 in-project additions
General time	KAUNO ontology era facet
Concrete time	Date resources in the project
Keyword	KAUNO ontology, 1034 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
Reference links	Link description resources in the project
Reviews	Review resources in the project
Cataloguer	Actor resources in the project

Literal Properties
Name, Alternative title, Textual description, Text sample

Table 3

Properties for abstract works stored in the database

Object Property	**Source of Values**
First publication	Boolean resource in the project
Original work	Link to original work if a translation
Language	LEXVO language ontology
Publisher	Publisher resources in the project
Publication year	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 Properties
Name, Subtitle, Number of pages, Complementary information about publication history

Table 4

Properties for concrete works stored in the database

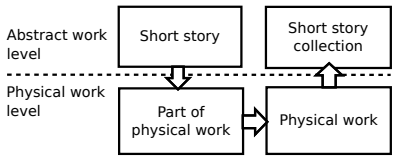

Fig. 1. Relationship between a short story and a short story collection in the BookSampo data model.

ing along with manual correction [3,8]. If a need for further separation arises, then one could just repeat a similar split procedure for the BookSampo data.

A minor problem in this model was how to represent short stories and their relationship with short story collections. Here, each short story has an abstract work level, which is "embodied" as a "part of a physical work". This "part of a physical work" is then included in a physical work, which in turn is the "embodiment" of the short story collection as an abstract work. This solution 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, belong 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.

The tables describing the fields of the abstract and physical works show that 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 5.

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. However, sometimes a particular resource actually doesn't correspond to an independently existing, understandable thing. This happens in cases where a relation has metadata of its own, such as when one wants to record the year in which 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 (blank node), where

Picture	
Literal:	Name, URL, Description
Book Cover	
Literal:	Name, URL, Description
Object:	Illustrator (Actor resources in the project), Keyword (KAUNO ontology, local keywords)
Keywords Belonging Together (e.g. suicide : justification)	
Literal:	Name
Object:	Keyword (KAUNO ontology, in-project additions)
Web Link	
Literal:	Name, Description, URL
Fictional Character	
Literal:	Name, Description
Object:	Keyword (KAUNO ontology, in-project additions), Personification of (Actors. This way, real persons and their fictious versions remain separate, yet linked)
Award	
Literal:	Name
Object:	Award Series (Award Series in the project), Award Year (Dates in the project)
Award Series	
Literal:	Name, Alternate name, Description
Object:	Keyword (KAUNO ontology, in-project additions)
Part of Series	
Literal:	Name, Number in Series
Object:	Series (Series in the project)
Series	
Literal:	Name, Description
Object:	Keyword (KAUNO ontology, in-project additions)
Literary School or Period	
Literal:	Name, Description
Object:	Concrete timespan (Dates in the project), Concrete place (KOKO-Place ontology, in-project additions), Keyword (KAUNO ontology, in-project additions)
Position of Trust	
Literal:	Name, Alternate name, Description
Object:	Keyword (KAUNO ontology, in-project additions)
Place	
Literal:	Name, Description, Latitude, Longitude
Object:	Larger Place (KOKO-Place ontology, additions)
Date	
Literal:	Name, Earliest Possible Start, Latest Possible Start, Earliest Possible End, Latest Possible End (ISO 8601 dates)

Table 5

Properties for secondary resources

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 the part number of 7. This caused problems because these auxiliary resources appeared to the indexers exactly like normal resources, 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 to the detriment of the indexers. Luckily, in practice the effects of this could be somewhat mitigated by training and annotation instructions.

5. Uses Cases for the BookSampo Dataset

The original use case for the BookSampo data is in the BookSampo portal, which provides an end-user view into the database. The dataset has also already been used in a context outside the original environment it was designed for, providing concrete examples on what can be done with the data. First, the dataset has been aligned with both the domain ontologies as well as collections of various museums, libraries, media organizations and archives, and published in the cross-domain cultural heritage portal CultureSampo[14] [7]. Second, On May 23, 2011, the major Finnish newspaper Helsingin Sanomat organized an open data hacking event, which utilized the BookSampo database through its 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. It also became possible to discover which themes were popular in fiction at various times. Other results combined BookSampo data with external grant data, showing for example what types of topics most likely receive grant funding or awards. New interactive applications were also created, showing which years were statistically similar from a publishing viewpoint, or locating the places associated with Finnish fiction on a map.

[14] Operational at http://www.kulttuurisampo.fi/

Acknowledgements

Thanks to Erkki Lounasvuori, Matti Sarmela, Jussi Kurki, Joeli Takala, Joonas Laitio, and many others. This research is part of the FinnONTO project 2003–2012, funded mainly by Tekes. The BookSampo project is funded by the Finnish Ministry of Education and Culture. The Finnish Cultural Foundation supported CultureSampo development.

References

[1] Ian Davis and David Galbraith. BIO: A vocabulary for biographical information. http://vocab.org/bio/0.1/.html.

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

[3] Thomas B. Hickey, Edward T. O'Neill, and Jenny Toves. Experiments with the IFLA functional requirements for bibliographic records (FRBR). *D-Lib Magazine*, 8(9), Sept. 2002.

[4] Eero Hyvönen. Developing and using a national cross-domain semantic web infrastructure. In Phillip Sheu, Heather Yu, C. V. Ramamoorthy, Arvind K. Joshi, and Lotfi A. Zadeh, editors, *Semantic Computing*. IEEE Wiley - IEEE Press, May 2010.

[5] Jussi Kurki and Eero Hyvönen. Collaborative metadata editor integrated with ontology services and faceted portals. In *1st Workshop on Ontology Repositories and Editors for the Semantic Web (ORES 2010), the Extended Semantic Web Conference ESWC 2010*. CEUR Workshop Proceedings, Vol-596, 2010.

[6] Eetu Mäkelä, Kaisa Hypén, and Eero Hyvönen. BookSampo—lessons learned in creating a semantic portal for fiction literature. In *Proceedings of ISWC-2011, Bonn, Germany*. Springer-Verlag, 2011.

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

[8] Jeremy Nelson and Alan Cleary. FRBRizing an e-library : Migrating from dublin core to FRBR and MODS. *code{4}lib*, (12), December 2010.

[9] Pat Riva, Martin Doerr, and Maja Zumer. FRBRoo: enabling a common view of information from memory institutions. In *World Library and Information Congress: 74th IFLA General Confrence and Council*, August 2008.

[10] Jarmo Saarti. *Aspects of Fictional Literature Content Description: Consistency of the Abstracts and Subject Indexing of Novels by Public Library Professionals and Client (in Finnish)*. PhD thesis, University of Oulu, Finland, 1999.

[11] Sami Serola and Pertti Vakkari. *Yleinen kirjasto kuntalaisten toimissa; Tutkimus kirjastojen hyödyistä kuntalaisten arkielämässä*. Finnish Ministry of Education and Culture, 2011.