Undefined 0 (0) 1
IOS Press

How to deal with massively heterogeneous cultural heritage data – lessons learned in CultureSampo

Eetu Mäkelä * and Eero Hyvönen and Tuukka Ruotsalo
Semantic Computing Research Group (SeCo), Aalto University, P.O. Box 15500, FI- 00076 Aalto, Finland

Abstract. This paper presents the CultureSampo system from the viewpoint of publishing heterogeneous linked data as a service. Discussed are the problems of converting legacy data into linked data, as well as the challenge of making the massively heterogeneous yet interlinked cultural heritage content interoperable on a semantic level. In the approach described, the data is published not only for human use, but also as intelligent services for other computer systems that can then provide interfaces of their own for the linked data. As a concrete use case of using CultureSampo as a service, the BookSampo system for publishing Finnish fiction literature on the semantic web is presented.

Keywords: Heterogeneous linked data, cultural heritage, services

1. Introduction

Cultural heritage is a field encompassing a wide range of content that varies drastically by type and properties, but is still semantically extremely richly interlinked. Currently, this content still mainly resides in closed databases, distributed nationally and internationally in different locations, and organized commonly by content type – separate databases are typically used for different contents, such as books, artifacts, videos, music etc. The organizations managing these databases are of different kinds, such as museums, libraries, archives, media companies, and web 2.0 sites. Moreover, different natural languages and cataloguing practices are used in different countries and organizations.

As an example, in Finland the Finnish National Gallery[1] holds a painting by the painter Akseli Gallen-Kallela, depicting a scene from the Finnish national epic, Kalevala, collected by Elias Lönnrot. The original poems are available as a database provided by the Finnish Literature Society[2]. There is also a separate database containing further information on each passage, actor and imaginary place of the epic poem. In addition, the National Biography[3] contains biographical information about Lönnrot and Gallen-Kallela and 6,000 other famous Finnish authorities, whose life stories may be mutually intermingled. At the same time, the Agricola-network[4] of Finnish historians holds a database of historical events, some pertaining to Kalevala, and others to Gallen-Kallela and Lönnrot. Yet further, the international Union List of Artist Names[5] by the Getty foundation contains structured information on other people Akseli Gallen-Kallela worked with as well as his roles in society. The video archives[6] of the Finnish Broadcasting Company contain videos related to all of these themes. In addition, all the sources contain lists of related places, times and content keywords which further link the material to mu-

*Corresponding author. E-mail: eetu.makela@tkk.fi.
[1]http://www.fng.fi/frontpage
[2]http://www.finlit.fi/index.php?lang=eng
[3]http://www.kansallisbiografia.fi/english/
[4]http://agricola.utu.fi/
[5]http://www.getty.edu/research/tools/vocabularies/ulan/
[6]http://yle.fi/elavaarkisto/

seum collection objects, photographs, historical buildings, and so on. As a final source of additional information, Wikipedia contains further peer-curated information on nearly all aspects of life, given that they are notable enough.

If the content could be extracted out of these closed repositories and integrated on a semantic level, it would be possible to start looking at culture as a whole instead of just gazing at it in thin slices, thus hopefully providing useful new functionalities and possibilities. As one of the main promises of the semantic web is easy integration of data from distributed heterogenous sources, cultural heritage has been found to be a nice test environment for evaluating the technology [20]. Previous semantic web portals for cultural heritage, such as [26,52], have tackled the problems of semantic interoperability and content integration usually focusing on only on few content types, such as artifacts or photos. If larger datasets and a wider variety of content types are considered, such as paintings, music, films, and books from different European countries in Europeana[7], then the level of semantic information is typically shallow and the promises of the semantic web cannot be fulfilled. The ultimate case of this are the general search engines for all kinds of text-based content anywhere, such as Google.

In contrast, the system "CultureSampo – Finnish Culture on the Semantic Web 2.0" [24,25], presents an approach and a demonstration application online[8], capable of dealing with virtually all kinds of cultural heritage content types (artifacts, books, art, poems, places, persons, historical events, narrative stories, music, photos, cultural heritage processes etc.), and at the same time using and benefiting from semantically rich descriptions, based on ontologies and metadata schemas. CultureSampo is a prototype system for integrating cultural content on a Finnish national level, using also international related contents, and has been in development from the year 2003. The portal was opened in public in 2008.

CultureSampo has evolved as a series of three major prototypes since 2005, extending the MuseumFinland[9] [21,27] system into a semantic portal of cultural heritage contents of virtually all kinds, a kind of national linked data memory on the web. The CultureSampo vision, business model, overall structure, and thematic views form an end-user perspective are presented in [25]. The first prototype in 2005, CultureSampo I, was an experiment in utilizing events and narrative structures as a semantical "glue" of cultural heritage content [32]. The next proof-of-concept version (CultureSampo II) explored the event-based approach further and used an event-based knowledge representation scheme as an underlying model for harmonizing heterogeneous cultural contents [47] (CultureSampo II). The current public version online [24,25] (CultureSampo III), represents still another different incarnation of CultureSampo, where Dublin Core like metadata schemas and the linked data approach were used for harmonizing the contents. Here events are used for representing narrative structures and as a semantic glue, but not for harmonizing contents.

This paper concentrates on the public version of CultureSampo (III). In contrast to our earlier publications [24,25], the focus here is especially on the content harmonization viewpoint and on using the system as a service. We discuss lessons learned during the long development process in content creation, content integration, user interface design, and using the system as a semantic web service for implementing other cultural heritage applications. In its current form on the web, CultureSampo integrates information from over thirty organizations, with a multitude of tens of different content types and metadata formats, and hundreds of thousands of individual content items. During the timeframe of the project, the system and its components have gone through multiple major revisions, through which a rich source of experience has been gathered on various aspects of semantic integration.

First discussed are the general types of integration needed, and the conceptual choices taken in various iterations of the CultureSampo system. A major hurdle in the project has been getting data out of the original databases and into RDF format. Besides technical and data quality problems, the project also had to deal with wider questions of how different organizations need different data production pipelines and supports in order to expose their data in RDF as effortlessly as possible, as well as ensuring the quality of that RDF. These issues are discussed in section 3. The harvested RDF is stored in a triple store, enriched semantically using RDFS-based semantics, and then indexed semantically for scalable search and other functionalities (section 4).

The whole effort of integration would be wasted if no new functionalities were made possible by it. Section 5 presents the functionalities developed for Cul-

[7] http://www.europeana.eu/

[8] The portal is available at http://www.kulttuurisampo.fi/ in three languages, but most content is in Finnish.

[9] http://www.museosuomi.fi/

tureSampo, as well as the conceptual and implementation choices that were necessary to develop these functionalities.

A major issue in creating the system was that the organizations taking part did not want their data to only flow into a centralized search system outside of their control, but wanted themselves to benefit from the added value inside their own web pages and systems. They wanted the provided functionality and the additional data, but wanted to provide their own views and user interfaces to the data, from their individual institutions viewpoint. In order to do this, in addition to intelligent user interfaces, CultureSampo needed to provide its functionality as extensible and flexible web service interfaces. These interfaces are described in section 6.

To further concretize things, the paper ends with a full example on a specific project making particularly extensive use of both the data production architecture as well as the web service interfaces of CultureSampo. This project is BookSampo[10], a joint project by the public libraries of Finland to create a portal for Finnish and Swedish language fiction published in Finland, as well as the authors of that fiction. The semantically annotated content includes roughly 60,000 items of fiction (novels, short stories, plays, etc.) with 30,000 parts, 20,000 authors, 10,000 fictive characters, and thousands of other content items of the literature world. The content was initially harvested from library databases and then completed and corrected manually by tens of librarians in a collaborative web 2.0 effort on the web, using the national FinnONTO ontology system [22], ONKI ontology services [55], and the SAHA 3 metadata editor [39].

In conclusion, the paper discusses the contributions of CultureSampo, and summarizes the lessons learned during our work.

2. Dealing with heterogeneous data

In an ideal case, data integration on the semantic web happens transparently and effortlessly based solely on its use of global URI identifiers and the graph nature of RDF that relates them to each other. However, in practice, many data sources still use their own schema and reference vocabularies which haven't been mapped to each other.

The task of integration then becomes a task of enumerating the required equivalencies and relations between the properties, classes and instances in the datasets. In practice, it is useful to distinguish between two different types of integration, as the means by which these are resolved differ extensively, based on the scope of the task and the importance of getting the mappings right [37,4].

First, schema integration deals with mapping the representations of objects in different datasets to each other or to a common ground, so that the structural differences of the data sources can be bridged. Second, mapping must be done between the individual resources themselves. Most commonly, the resources that must be mapped are those referenced as property values in the original datasets, such as authors that come from different authority files, keywords originating from different reference vocabularies, or the identifiers of places stored in external geographical databases.

To shed light on the relative scopes of these mapping tasks, analyzing the data in the currently published CultureSampo portal as an example, there are about 200 truly semantically different properties among the 30 or so different content types of the portal. The number of primary content items on the other hand is a little over 600,000, while the number of resources referenced from these is a little over 6 million. On the other hand, in a test data set of a little over 4.1 billion triples crawled from the web and the Linked Open Data cloud, there are a total of some 350 million distinct resources, 1.3 million currently distinct classes in use, while the number of distinct properties in use is just less than 280,000, though probably a vast majority of all of these would have equivalencies between each other.

In the following, these two aspects of data integration will be discussed in turn, along with the solutions created in the CultureSampo system.

2.1. Schema integration

Schema integration most commonly involves going through each object type in each original data source, evaluating correspondences in types, properties and conceptual models between the data sources. Compared to the mapping of the individual resources between the databases, the task of schema mapping is usually both more important for the overall usability of the data, as well as of a much more manageable size. Therefore for CultureSampo, these mappings were done, or at least inspected by hand.

[10] http://wiki.kirjastot.fi/index.php/Projekti_Kaunokirjallisuushanke

In the course of the CultureSampo project, two different approaches to schema integration were tried. First, in CultureSampo II, the project tried mapping the schemas to a more primitive event-based representation, but when this ran into problems, the properties were eventually mapped using more traditional property mappings across schemas.

2.1.1. An event-based approach

The first approach to schema integration in CultureSampo II was to use events and thematic roles as a harmonizing representation format underlying the heterogeneous data [47,53]. As an example, consider the following metadata about a painting and a person (prefix definitions omitted for brevity):

```
cs:Kullervo_departs_for_war
  cs:painter person:A.Gallen-Kallela ;
  dc:date time:1901 ;
  cs:placeOfCreation place:Helsinki .

person:A.Gallen-Kallela
  cs:placeOfDeath place:Stockholm ;
  cs:timeOfDeath time:1931 .
```

Using mapping rules, the following corresponding event descriptions were generated:

```
cs:painting_event_45
  rdf:type onto:painting_act ;
  e:actor person:A.Gallen-Kallela ;
  e:target cs:Kullervo_departs_for_war ;
  e:time time:1901 ;
  e:place place:Helsinki .

cs:death_event_41
  rdf:type onto:death ;
  e:target person:A.Gallen-Kallela ;
  e:time time:1931 ;
  e:place place:Stockholm .
```

In the selected formalism, the number of relational properties drops to the six thematic roles of actor, target, instrument, goal, place and time, with the different event types being related to each other through an event ontology. Quality properties still stay the same, so that for example "cs:Kullervo_departs_for_war cs:technique cs:oil_on_canvas" would remain unchanged. What would be added, however, is a relation relating the property "cs:technique" to the concept of technique in an ontology. This way, as for the event types, the task of mapping could be pushed to the more well-defined space of class ontologies, that could then, for example, define that both painting and writing are subclasses of creation.

The benefit of the selected formalism was also that it allowed the formulation of clear cut rules for evaluating how different properties were mapped to the model. Having a simple, well-formulated static model as one mapping partner made the mapping task considerably easier.

Aside from providing a simple common target model for mapping, it was also envisioned that reasoning, user interface generation, and also user understanding of the schema would benefit from the drastic reduction in the number of properties offered by the approach. However, while the model did work sufficiently well as an underlying data model for reasoning and recommendation [48], it turned out that it actually created problems on the user interface level, as explained in the following.

For this version of the CultureSampo system [28], a view-based search interface [13,41] was prepared that used the event schema directly, i.e. the user could constrain material by "event type", "event location", "actor", "target", and "event time". However, when user tests were conducted on this prototype, the event-based views were criticized as unintuitive [30]. Based on this end-user evaluation, as well as interviews with personnel from organizations doing indexing for CultureSampo, it became apparent that while events may be a good base for tying content together and for reasoning, they are not intuitive to end-users.

Bringing events to the fore, the approach fractured and distributed the metadata of the original primary objects. This meant that traditional and well-understood attribute-value pair visualizations and queries could no longer be applied to the original objects. For example, to search for objects created by Akseli Gallen-Kallela, one could no longer just search for objects that had the "painter" property set to "Akseli Gallen-Kallela". One had to select "painting" as "event type", "Akseli Gallen-Kallela" as "actor" and select the "target" slot as the result sought.

In showing items, complex visualization were needed that placed them in relation to all the events that touched them, instead of being able to simply list the data as attribute-value pairs. These visualizations in turn were considered both by users and annotators as vastly less clear and usable than the original primary object-oriented metadata. Thus, at least for the user interface, the approach to schema integration in CultureSampo had to be rethought.

2.1.2. Traditional integration

In the end, the problem was resolved by merely organizing the properties in hierarchies using traditional means [33], and by solving discrepancies between conceptual models by domain specific mappings.

In general, the task of mapping properties and classes by hand between data sources turned out to be easy also in CultureSampo's heterogeneous environment. The twenty or so classes could simply be gone through by hand. The organization of the two hundred or so properties required a little more methodology.

Mainly, the differences in related properties between data sources were matters relating to level of generality. For example, one museum collection might use a general "place of creation" property, while another uses the more distinct "place of manufacture". In a collection of paintings on the other hand, these properties might be "place of painting" versus "place of creation", which could then be related to the former pair through their common range of "location".

In analyzing the properties, it seemed that the properties could mainly be grouped together through two distinct axes defined in turn by their domains and their ranges. First, the material was analyzed by range. Here, the basic ranges of location, time and person or organization were particularly prevalent and useful in categorizing the properties in simple understandable ways. In addition, the datatype properties could be categorized into e.g. properties dealing with object identification and descriptive properties. After this, the properties left over were grouped based on the object types in which they were found, such as relating all properties related to music or museum collection objects to each other.

Sometimes, the problems were however more thorny. For example, the schema for the Finnish Museums Online[11] system, an aggregator service in itself, contains the field "place of acquirement/place of discovery", which irrevocably combines these two fields found separate in other collections. Here, while the obvious solution would be to define this compound property as a super-property of both distinct properties, in the end we decided to do this in reverse as a matter of practicality, by making the compound property an *uncertain* sub-property of both parts. The rationale behind this was that while we could not exactly determine which property was being meant in the case of e.g. objects with "place of acquirement/place of discovery Helsinki", users would probably still like to see them while searching for e.g. objects discovered in Helsinki, as long as we showed alongside the items the proviso that the match was uncertain.

Also of note was that properties with the same name did not always mean the same thing. The property "color" in a museum database usually describes the coloring of the objects, while in a particular photography database it is a binary predicate with options "color" and "monochrome". Here again, the analysis by range proved useful.

As regards major differences in conceptual modeling between source materials, these were mostly discovered in relation to events, temporal entities and the coding of geographical location.

In the case of events, nothing is currently done to map between different conceptual models in CultureSampo relating to them. This is because, based on our prior experience as also reported here, the choice of modeling or not modeling a particular thing as an event is a choice that is decided based on how that thing is best understood semantically by a user. These choices also seemed consistent between different data sources, so that for example birth and death dates were uniformly annotated to the person objects they were tied to, while exhibitions, historical happenings and so on were modeled as events. This sensibility seemed also to hold when analyzing the matter from a user interface and linking perspective, so that for example a user viewing a timeline of historical events in the late 19th century would want to see any artists active at that time as artists (and not as artist birth and death events), while events not similarly primarily tied to another semantic object would of course be shown individually. One single exception to this was the CIDOC-CRM[12] [12] -based system used in the Finnish National Gallery. This is because CIDOC-CRM itself is an interchange format that is based primarily on events, similarly to our earlier approach. In practice in the FNG user interface, these event details were hidden and translated back to attributes, which again confirmed the hypothesis that the event formalism is not the primary way through which people want to view such attributes.

For geo-coordinates, most common differences were in the projections used. Besides such encoding differences, often also different object models were found, such that one source would say (using an object property) that a building was "located at" an object of the type "point", that had geo-coordinate properties, while others pinned the geo-coordinate properties directly to the building. In cases such as these, it was often enough

[11] http://suomenmuseotonline.fi/en

[12] http://www.cidoc-crm.org/

to just select one object model and map all the data to that.

The subject of differences in conceptual modeling relating to temporal entities was more interesting. Here, besides simple differences in encoding, even an identical annotation across different content types proved to often encode subtly semantically different information. As an example, consider the statement that something has a "time of creation" of "1830–1850". Now, for e.g. a building, this probably means that it was built either continually or in parts for the whole duration. For a painting it would probably represent uncertainty of dating, with the actual time-span for creation not exceeding a couple of years at most. Finally, for an industrial item, it could mean either that that particular object was made sometime between 1830 to 1850, or that the particular model the individual object represents was manufactured during that timeframe.

To be able to consolidate all these different conceptions of fuzzily defined temporal entities, the final temporal schema of CultureSampo operates on a fuzzy temporal model [56], where each temporal entity is constrained by defining the earliest and latest possible start and end dates, as well as possibly further definitions on minimum and/or maximum duration for the event based on content-specific rules. For an exhaustive discussion on the semantics of this model as well as what can be done with it, see [34].

2.2. Reference resource integration

The same problems of integration that apply to properties also apply to the values of the properties, i.e. different collections may use different vocabularies, such as one designating an item as "man-made" while another uses "crafted by hand". Also the annotation level of granularity may differ, such as one collection making a distinction between a chalice and a goblet, while another would classify them both just as drinking vessels. In CultureSampo, luckily, these problems were much diminished, because Finnish libraries and Museums have a long tradition of drafting and making use of common vocabularies. However, all these vocabularies were still special to a single field such as fiction literature as opposed to museum artifacts, or works of fine art, without links between the fields.

To solve this problem, the project leveraged and relied on the work done in the wider FinnONTO project [22], 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 special field ontologies as possible.

As stated, Finland has a long tradition of utilizing common vocabularies, particularly in the cultural heritage domain. To leverage this ready resource and also to make mapping collections indexed with those legacy thesauri easy, most of the ontologies in the KOKO infrastructure are light-weight ontologies transformed from these original vocabularies. This is done using a process where first the thesauri are translated into SKOS or OWL automatically, after which their subsumption hierarchies are checked and mapped to those already in the infrastructure, most importantly the Finnish national upper ontology YSO[13], which forms the backbone of the infrastructure. The current constituents of the KOKO ontology infrastructure are listed in table 1.

In addition to these class ontologies, the infrastructure also contains further instance registries, such as a geographical registry of 800,000 places in Finland and a further 5 million abroad, a spatiotemporal ontology of Finnish counties 1865–2007 [36], and an actor ontology of about a million persons and organizations.

If a source material to be integrated references a thesaurus or categorization not included in the core, quality controlled KOKO infrastructure, that reference vocabulary is converted into RDF and mapped to KOKO concepts automatically where possible. This allows for maintaining the quality of subsumption inference while still linking the material to each other as much as possible. In CultureSampo, such resources used are for example the Iconclass[14] classification for describing subject material in fine arts, as well as the HKLJ[15] and YKL[16] literature catalogue classifications used in Finnish library systems.

3. Data pipeline

Having settled on a model for how the data coming in from different data sources would be integrated, the following task was to discover how best to ensure the flow of high quality data from the 30 or so orga-

[13] http://www.yso.fi/onki3/en/overview/yso
[14] http://www.iconclass.nl/
[15] http://hklj.kirjastot.fi/en-GB/?PrevLang=fi
[16] http://ykl.kirjastot.fi/en-GB/?PrevLang=fi

Table 1
KOKO ontology constituents

Name	Ontology domain	Underlying thesauri	Size	Languages	Maintaining organization
YSO	Upper ontology, general domain	General Thesaurus YSA / Allärs	23700	Finnish, Swedish, English	National Library and Åbo Academy
MAO	Museum domain	Thesaurus of the Museum Domain MASA	6800	Finnish	National Board of Antiquities
TAO	Applied arts	Thesaurus of Applied Arts	2600	Finnish	University of Eastern Finland and Library of Aalto University
MUSO	Music	Thesaurus of Music MUSA / CILLA	1000	Finnish, Swedish	National Library
KAUNO	Literature subjects	Thesaurus of Literature Kaunokki / Bella	4900	Finnish, Swedish	Finnish Public Libraries
VALO	Photography	Thesaurus of Photography Literature, Thesaurus of Photography Technology	1900	Finnish	Finnish Museum of Photography
KITO	Literature research	Thesaurus of Literature Research	900	Finnish, English	Finnish Literature Society
KULO	Culture research	Thesaurus of Folk Culture Studies	1600	Finnish, English	Finnish Literature Society
KTO	Linguistics	Thesaurus of Linguistics	1000	Finnish, English	Research Institute for Languages in Finland
POIO	Points of Interest	TGN, Geonames, OpenStreetMap, SUO	1000	Finnish, English	Various
AFO	Agriculture, forestry	Agriforest Thesaurus	5500	Finnish, English	Viikki Science Library
MERO	Seafaring, shipping	Thesaurus of Seafaring	1400	Finnish	Finnish Transport Agency
JUHO	Public government	Thesaurus of Finnish Government, VNAS	6400	Finnish	Ministry of Finance
PUHO	Defense	Thesaurus of Defence Administration	2000	Finnish	Finnish Defence Forces
TERO	Health promotion	YSA, HPMulti, MeSH, Stameta	22000	Finnish, Swedish, English	Various
Total			82700		

nizations participating in the content production of the CultureSampo system.

In the wider scope of things, the FinnONTO project aims at bringing semantic indexing to the original cataloguing and indexing systems. To this end, the architecture provides, for example, ready to use web widgets for picking concepts from the KOKO ontologies that can be integrated into an existing browser-based content management system with merely a few lines of JavaScript [43]. However, at the time of gathering the material for the CultureSampo project, these capabilities were not yet ready. In addition, most of the participating organizations also were not yet ready to take the leap into full semantic indexing.

Thus, in most cases, the flow of data into CultureSampo had to be organized around transformers that took the original non-semantically indexed data and mapped it into RDF and enriched it. Also, even those few organizations that could adopt an RDF-based metadata editor, such as the FinnONTO-created SAHA [39], had to first transform their legacy data into RDF.

As stated, the approach taken in CultureSampo was to map schema representations manually, and use the linked ontology infrastructure of KOKO to map referenced values. While the former did prove easy, the latter proved problematic as a result of sloppy indexing practices in the original data sources.

A major source of problems here was that in the absolute majority of legacy systems, all field values were entered as text, even those that in an ontological environment would be modeled as objects. In mapping literature, this is known as the impedance mismatch between traditional databases and the object-oriented world of the semantic web [45]. While theoretically it should be easy to just take the textual values selected from a thesaurus and match them to the ontological entities in the corresponding ontology, in practice problems abound. For example, typically a significant minority of the values contain misspellings. Also, even when the use of some thesaurus was mandated

for a particular field, there was no guarantee that the values actually came from that thesaurus. In addition, the means by which multiple selections in a field were recorded could vary between items and indexers, such as one indexer separating them by spaces, while another used commas or semicolons.

Another problem was that in many cases, the users of the original databases had found them lacking and created ad hoc fixes that went against the general semantics of the fields. A particularly common example was when a work of art or museum item consisted of multiple parts. For example, in the National Bureau of Antiquities database, the semantics of the technique field is that it should contain a comma separated list of techniques taken from the MASA thesaurus that describe how the item was made. However, multiple items in the database contain compound textual explanations, such as "hood, made by hand, horn, made by a silversmith, base, factory-made".

Similar problems are evident also in the Finnish National Gallery content management system, which is based on CIDOC-CRM [12]. As one example, CIDOC-CRM only defines a single "consists of" field for describing the material an object is made of. However, in works of fine art, it is important to distinguish between background canvas material and the foreground material. In the FNG database, this has been achieved by placing both designations in "consists of" fields, but declaring that the one of the values comes from a fictional "foreground" thesaurus and the other from an equally fictional "background" thesaurus.

Quite commonly, these ad hoc fixes also removed the semantic separation generally maintained between fields. For example, in the Finnish Museums Online[17] dataset, if a "part" field appears, it must be assumed that further description fields following it relate to a particular part and not to the whole, until another "part" declaration. Thus, in such cases, one cannot simply take a field such as "technique" at face value in isolation, as the semantics of that field now depend on other properties and possibly also the order of those properties in the data.

In some materials, this sort of pattern was endemic. For example, in the HelMet[18] cataloguing system of the Helsinki metropolitan area public libraries, many fields contained seemingly extraneous commas, colons, slashes or quotes at the start or end. It turned out that these were present so that when the content of the fields was concatenated in a particular order, the resulting presentation matched that used in the library user interface.

The task here was thus to create a pipeline architecture capable of 1) converting legacy data into RDF and 2) mapping the resulting schemas and resources into the global CultureSampo pool, all the while being able to handle the noise and errors inherent in the original data. In addition, a need for semantic enrichment of some of the data sources was identified. For example, some data sources didn't contain separate content keywords that could be mapped to ontological concepts, but did have textual descriptions of the content from which such key concepts could be mined.

At first, monolithic converters were created for each material separately. However, it was soon discovered that this resulted in large-scale duplication of common processing steps, as well as led to differences in how each data set was mapped. Also, when discovering a new problem in the mapping or in the original data, any changes and exceptional processing needed to be done into the program code of each converter program and the whole process had to be run again from the start.

The challenge then became to minimize implementation costs by designing an architecture that contained as many reusable parts and common processing phases as possible, despite the data sources varying so much in both content and original format. Also, to make understanding of the functioning of the pipeline easier, as much separation as possible was to be maintained between the processing steps. In the end, this led to a four part pipeline architecture. In its principles, this architecture can be seen as a refinement of the one previously developed and tested in the MuseumFinland project [26], as relates to how it deals with different input formats and differences and exceptions in the reference vocabularies used in the sources. However, for CultureSampo, the approach had to be refined and altered in many ways to be able to deal with the vast increase in heterogeneity of content types, schemas and vocabularies.

3.1. Conversion to RDF

As a first processing step, the material was converted from its original format into RDF. However, instead of directly following best practices for encoding such information in RDF, this processing step aimed at reproducing the structure of the original data model in RDF as much as possible. This meant, for example,

[17] http://suomenmuseotonline.fi/en
[18] http://www.helmet.fi/search S9/

that references to common vocabulary terms or other objects were copied as datatype properties instead of being mapped to resources.

The purpose of this step was mainly to isolate the problems of dealing with a legacy format away from all the semantic processing and model transformation, which could then be done solely in RDF. This proved useful for example when one source switched publication format (but not the essential source data model) midstream. All changes could be done into the first processing step, with the rest continuing to operate transparently on the RDF version of the data.

Also, not doing semantic processing related to data model issues at this first step, allowed a common converter to be used even for sources with semantically different conceptual models, as long as they shared just the original source format, such as XML, Excel sheets or SQL.

3.2. Model harmonization

The second processing step was then to convert the conceptual model of the original data to conform with semantic web best practices. Most usually, this meant converting textual references to common vocabulary concepts and individuals into references to shared URI resources.

However, in many collections, it was also the case that some of the individuals referenced *were* modeled as objects. However, they were not present only once, identified by a common identifier, but multiple times as anonymous resources whose information was copied across all primary object annotations that referenced them. This again was mostly the result of these collections and their serialization formats being organized around a single core content type. For example, in the XML serialization of the Finnish National Gallery database, all primary artifact objects contained an inlined anonymous person object describing the creator of the artifact, even when these creators appeared numerous times in the data. Even more prevalent this was in regard to structured temporal objects in the data sets, such as years of publication, opening times, exhibitions where an object had been shown and so on. At this stage of processing in CultureSampo, such multiple instances of the same individual or concept were simply collapsed to a single common URI.

While earlier attempts at the pipeline had at this stage tried to find the common resource URIs to be referenced by label in the KOKO ontologies, the final version of the data pipeline architecture operates in data space isolation, simply creating the resources and their URIs in a custom, data source specific namespace based on their textual labels. The reasons for this are twofold. First, it was desired that no information was lost in the processing even if mappings were not found. Previously, any mistyped keywords or malformed fields were just lost at this stage, with a rerun needed if any corrections to processing were made. The second, albeit linked reason was a desire to push mapping resolution to a separate, uniform processing step. Having everything, even the errors simply mapped to resources allowed a later mapping stage to be iterated multiple times and improved without simultaneously having to run all data conversions again. This separation of concerns was enforced also regarding all the properties and classes used, which were also at this stage manufactured into a custom namespace in an equivalent manner to the resources.

While this phase had to be implemented for each conceptual model separately, it could still use common functionality for example in separating multiple values from textual keyword fields. This allowed all data production pipelines to benefit from any centralized improvements to these algorithms.

3.3. Schema and value mapping and semantic enrichment

Dropping any mapping resolution out of the previous stages allowed the mappings to be done centrally and iteratively in the global CultureSampo data space. Actually, this task became one with the general task of mapping different RDF materials to each other, and could use any of the readily available ontology and instance mapping tools [33,7] for doing just that.

These mappings were then stored in RDF using the OWL, RDFS and SKOS equivalency, distinction and subsumption properties, to be resolved in the triple store engine at runtime. This also meant that any erroneous mappings could be undone easily after the fact by just removing the RDF triple specifying the bad mapping. Also, because all references in the original materials were converted into resources equally, this meant that no information was lost at any point in the processing, and when a more advanced mapping algorithm became available or human resources could be allocated to go through the outliers, the mappings could be improved iteratively.

Any semantic enrichment done to the materials is also done in this global data space. This ensures that, for example, when searching for concepts from textual

descriptions, the algorithms have a maximal amount of content available from which to draw matches.

The tasks of verifying and improving the resource mappings generated, as well as verifying automatic enrichments, can be done in the CultureSampo ecosystem through the SAHA metadata editor [39] created in the FinnONTO project, which has special support for going through annotations marked as suspect. The marking of such annotations can either be done originally in the enrichment process, or at a later date by utilizing heuristic or schema-based quality assessment rules.

For this latter task, the FinnONTO architecture contains the semantic content validation service VERA[19]. The output that VERA produces is not a list of errors per se, but rather a list of possible problems that an expert user can assess, and modify the schema or data as needed. The report also contains general statistics about the data, such as language definition usage, so it can also be used for a general analysis instead of validation.

4. Inferring triple-store

Our final choice of semantic integration by mapping properties and resources in RDF placed certain requirements on the inference capabilities of the triple-store used to query the contents. Specifically, the triple-store had to support easy and efficient resolution of both equivalency as well as subsumption relations, as those were the primary means used to map content.

In fact, in the custom triple-store implemented for CultureSampo, both of these are done transparently. As an example, a query for "?s rdfs:label ?o" would return also all skos:prefLabel and skos:altLabel triples, as well as any custom schema properties marked as equivalent to any of these. A query for "?s rdf:type foaf:Agent" on the other hand returns also instances of all the sub-classes of foaf:Agent. For ease in additional processing, a unified view to the data is also provided, where all URIs in an equivalency set in the source are replaced with a single canonical version. This way, anyone processing the results of such inferred queries need not themselves repeat the equivalency calculation.

The mapping principles adopted for CultureSampo match those used in the linked open data movement[20] [5,14]. It is interesting to note how the CultureSampo engine handles LOD data. In a test run, the system could load and serve a 4.1 billion triple crawl of LOD and web data. Here, a total of 11 million equivalency sets were discovered, touching 25 million resources out of a total 350 million.

In addition to subsumption and equivalency inference, the triple-store of CultureSampo also includes support for quickly discovering all location resources annotated with geo-coordinates inside a specified bounding-box, as well as all other resources related to those locations. The same is done for the fuzzy temporal entity resources and further resources related to them. These are all functionalities that were needed in the various user interfaces of the portal. Similarly, efficient text search is provided for searching 1) objects by their labels, 2) objects by their literal attributes and 3) objects by the labels associated with their object attributes. The last index is used in the general text search interface of CultureSampo, so that one can for instance query by the string "Pyhäjärvi" and be quickly returned all objects that relate to any of the 50 or so lake Pyhäjärvis of Finland.

5. User Interface Elements

Having managed to get the content into RDF and mapped to common vocabularies and schemas, the task of building user interfaces began. Here, the core question to be answered was what added value could be gained from the combined and heterogeneous nature of the sources. In this section, added value for the human end-user is considered first from generic semantic browsing and search perspectives. After this, CultureSampo's functionalities in providing different thematic views to a single semantic RDF service are briefly overviewed.

5.1. Semantic browsing

Already the basic semantic browsing that was made possible by joining and mapping the disparate data sources into a unified whole was rewarding. It is now possible for an end-user to browse culture by topics, facets, and via semantic associative links, without regard for content type or originating institution. Using the versatile and semantically rich RDF content, one can, for example, jump from a particular rune of the epic Kalevala to a painting depicting the events of that rune, and from there on to other works on Kalevala

[19] http://www.seco.tkk.fi/services/vera/
[20] http://linkeddata.org/

topics by the same painter, or to his biography, or visual depictions of his social and historical spheres of influence. Taking another turn from the original rune may lead her to the original poem collections from which the Kalevala was pieced together, the collectors of those poems, and the places where they were collected from, with associated historical photographs, and so on.

This kind of browsing resembles tabulators used for Linked Data [14]. However, in the case of CultureSampo, the associative links are not necessarily explicit RDF links as in Linked Data browsers, but may be created online by special recommendation rules expressed as SPARQL queries. The desired logic behind recommendations can be extracted from e.g. curators in museums, and be changed easily, too, without touching the RDF graph.

5.2. The search and organize user interface concept

In relation to search interfaces, the work done in CultureSampo for search on heterogeneous data yielded not only technical solutions, but an argument for shifting focus in semantic search from items themselves to using them as lenses to wider topics.

Traditionally, Internet search has been about finding a document or documents that answer the question posed by the searcher. Semantic Web search systems have mostly also held this viewpoint [17], using properties and concepts in domain ontologies to locate search objects annotated with them. For semantically annotated content analogous to text documents, this works adequately, but for qualitatively different material, it creates problems. To understand why, one must take a step back to look at information needs.

Classifications of information needs [59,3,10,8,31, 2] agree that there is a major partition between look-up queries like "For my meal, I need a *white wine* with a *spicy flavor*" and more general information needs such as "tell me all about *spicy white wines*". The former focuses on selecting, fact finding, and question answering, while the latter deals with the more general objective of learning and investigation, containing in addition to searching also tasks such as comparison, interpretation, aggregation, analysis, synthesis, and discovery [40]. Depending on the domain, at least a significant part (22% [9]), or even the majority (70% [58], 67% [8]) of inquiries for information relate to learning as opposed to spot queries.

Despite this, search research has only recently begun to move to this expanded domain, termed exploratory search [40]. We propose that a major reason for this is that as long as the information is encoded only inside documents, learning and investigation searches are adequately catered for by the same functionality as fact finding, i.e. locating all matching documents and then perusing each for relevant data [31].

For semantically annotated content other than information documents, the situation is different. Often the useful information is not the object itself, but the relation between the object and the ontological resources associated with it. Now, for question answering such as what wine to have with a particular food, the answer is still a particular object with particular characteristics, and the old paradigm still works. For the more general type of queries, on the other hand, typical Semantic Web object databases fall short, as they contain no singular exposition about, e.g. "French spirits".

However, if looked at from another perspective, the data contains ample information to answer someone wanting to know about French liquors. It is merely encoded differently, distributed across the multiple object annotations and ontologies. To pull this information out, one must move the focus from individual items to the set of objects with particular properties as a whole, and even further. What one actually wants is to look at the combination of the domain concepts "French" and "spirits" through the lens of the items.

Actually, if an interface capable of such can be created, the pieced nature of the information becomes an advantage, as the pieces can be combined to shed light on a much wider variety of topics than anyone could write an explanatory article on. This capability is even further enhanced if the database contains material of multiple different kinds. For example in the cultural heritage domain, with suitable material, one could learn not only about 19th century Finnish crafts, 19th century Finnish paintings etc., but actually of the 19th century Finland as a whole.

Based on this analysis, we argue that to support exploratory search tasks, Semantic Web application designers need to shift focus from finding objects to the creation of structured, domain-centric presentations based on those objects.

5.2.1. Looking at culture through its products

Luckily for interface designers in the cultural heritage domain, there is already a real world counterpart for this functionality to take inspiration from. What is wanted is very similar to how exhibitions in real-world museums function, presenting a particular temporally,

spatially and functionally constrained aspect of culture through its objects and art. As such parallels are an excellent cue for understanding the structure of an information presentation, it was decided to make as much use of this as possible when designing the interface for the CultureSampo portal.

The idea of the CultureSampo search and explore interface is to let users create virtual exhibitions that mimic the way real museums are organized, containing themed exhibition rooms of items and displays that together, through the objects, tell the story of a particular subject. The implemented system combines an exhibition specification interface based on view-based query constraining with multiple visualizations grouping the items according to domain facets the user is interested in. In the following, both of these components will be discussed further in turn.

5.2.2. Specifying the desired exhibition

The CultureSampo portal is aimed at the general public. Therefore, the exhibition generation interface had to be as easy to use and understand as possible, while still allowing for a wide variety of different presentations to be generated. To accomplish these goals, we first set the parameters for what kind of exhibition definitions had to be possible. Analyzing real exhibitions, a general but verbally understandable structural pattern for describing them was created, on which the interface could be built. The pattern, with each part optional, is:

> Tell me about *item type*
> related by *role* to *domain concept* [and . . .]
> organized by *classification*+*role* [and . . .].

While constrained and procedurally structured, this pattern still allows for a wide range of exhibitions to be specified, from e.g. "Tell me about weapons" to "Tell me about everything related to 19th century Finland and agriculture, organized by item type and purpose of use" and "Tell me about toys manufactured in China, organized by time of manufacture and place of use". Figure 1 shows this in the actual interface. On the left are the exhibition specification controls, laid out to directly reflect the developed narrative structure.

5.2.3. Domain-centric view-based constraining

For the selector components used to fill this pattern, we looked to the recently popularized [18,51, 16,44] paradigm of view-based search—also known as faceted search/browsing—combined with semantic autocompletion [23]. View-based search is based on organizing the search data into multiple categorizing views and then picking categories as constraints from the view facets, and has already shown good promise for fulfilling learning type search needs [57].

For the particular needs of the search and organize interface of CultureSampo, the paradigm has a number of user benefits [41]. First, because the collection is visualized along different categorizations, the user is immediately familiarized with its contents and the way they are organized. Functionally, the user gets information on what the possible constraints are and how selecting them will affect the result set. Second, the multiple viewpoints allow the user to start constraining from the perspective most familiar to them. Finally, this visualization already intuitively shows the wider context in which the result set lays, thereby contributing to the users ability to answer questions of the result set as a whole, and not just of individual item.

In addition to interface benefits, the paradigm fits Semantic Web data well. The rich metadata in semantic databases is just the sort of multifaceted data whose exploration the paradigm supports. Also, because the metadata values are resources organized in ontological hierarchies, they provide an excellent basis for creating usable, well-structured categorizing views.

Traditionally in Semantic Web view-based search systems views have been formed by selecting a property, such as "place of manufacture", and enumerating all the values of that property as selections. When using semantically heterogeneous content, this causes problems, since the different content types may be categorized along different facets. For example, in CultureSampo there are properties such as "mentioned place" (poem) and "depicted place" (painting, photograph) [16], that do not make sense with respect artifacts such as chairs. Fortunately, our move from the objects to the domain concepts presented us with a natural solution, the novel variation of domain-centric view-based search. Here, the properties are relegated to a secondary role, and the views were built based on a set of topical domain ontologies instead of the ontological ranges of those properties. In CultureSampo, we ended up with nine views: object types, places, times, actors, events, styles, materials, techniques, and museum collections, with attached properties such as place of manufacture, depicted place and place of birth.

In the former CultureSampo II [28], the properties were discarded from the user interface completely, and the views selected all items related in any way to the domain concepts (e.g. show anything related in any

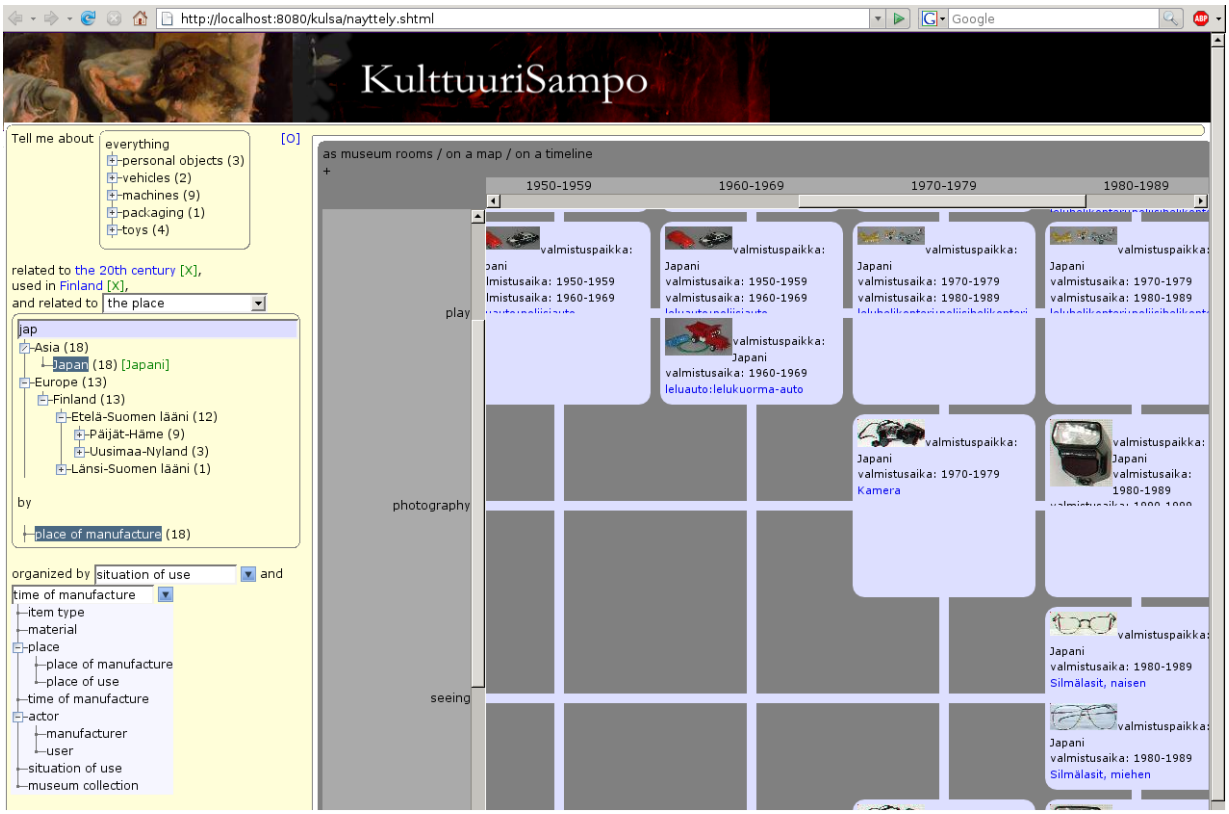

Fig. 1. The CultureSampo user interface, with important elements manually translated into English. The exhibition specification interface is located on the left, while the exhibition itself is visualized on the right. Showing is an exhibition on the types of items Japan exported to Finland in different parts of the 20th century.

way to Poland). However, without any reference to the properties, the users were lost as to what a selection did and why any particular item was included in the result set. In addition, the expressive power of the interface diminished, as one could no longer e.g. search for items made in Japan but used in Europe.

These problems were solved by two measures. First, in the presentation, for each item an explanation is included of the property-concept relationships that places that item in the result set. Second, the properties were brought back to the views, but in a different form, shown in the place facet of Figure 1. Now, a view consists of two selectors: one for selecting the domain concept and another for limiting based on the role (property) that the concept has in relation to the search items. Here, the user is free to search both with and without specifying a role, actually increasing the expressiveness of the view-based search paradigm.

In CultureSampo the views are not all constantly visible. This is because here they are used as selectors in the context of a larger pattern, which we wanted to emphasize. Showing many views at once by default would have cluttered the screen, reducing intuitive grasp of the interface. Instead, two views are visible by default, one static for constraining by item type, and another for constraining by a domain concept. The domain view visible in any given moment is selected from a drop-down menu. In addition, power-users can also bring up further concurrent views.

5.2.4. View-independent semantic autocompletion

The multiple views in the view-based search paradigm make it easy for users to browse their options. However, for users knowing precisely what they want, a shortcut and a single point of entry is desired. In our system, this is accomplished by a semantic autocompletion [23] component, shown in Figure 2.

Here, the user merely types in what they are looking for, and the system instantly responds with matching keywords to be used as possible constraints. These are both annotations directly related to the items, as well as matching selections in any of the facets. If the keyword typed gives sufficient specificity for the user, it isn't

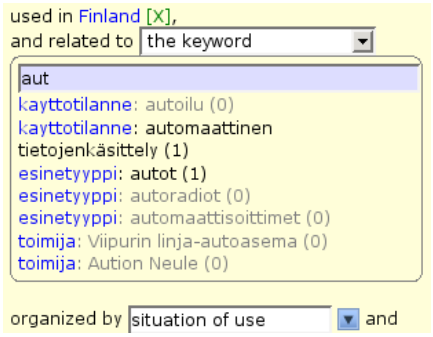

Fig. 2. The view-independent semantic autocompletion component of the CultureSampo exhibition specification interface.

even necessary to make any further selections, as the query state is also instantly updated, using the union of the matches as a constraint. This makes it possible for a user to interact with the system in a more experimenting way, typing in a keyword that pops into their mind and immediately seeing if the portal contains any related material, as well as what kind of exhibition it generates.

These keyword-search derived constraints can also be combined with those selected from domain views. For further supporting in-between user behaviours, all the domain views internally support a different form of semantic autocompletion, with the results shown directly in their hierarchical view context. This functionality is depicted in the place facet of Figure 1.

5.2.5. *Visualizing the exhibition*

As the user makes choices constraining the material, also the exhibition view is updated. Here, the primary association sought is of a typical museum, with themed floors and rooms of exhibits, combined with custom presentations.

For the museum room visualization, the same categorizing view structures used for selection are utilized. The idea is simply to project the items in the result set onto a two dimensional matrix whose rows and columns are comprised of a flattened list of concepts in the two domain facets chosen for organization. This way, each cell in the matrix corresponds to room combining two themes, such as “18th century agriculture”, followed in one dimension by “19th century agriculture”, and “18th century hunting” on the other. This matrix is then visualized, either as is for a single-floor museum complex view depicted in Figure 1 or row by row, for a more traditional floor and room museum plan. While the latter plan allows for eliminating empty rooms on a floor by floor basis thereby optimizing display area, the single-floor view allows one to also see more large-scale structural changes. In Figure 1, for example, one can see how in 1950-1970 most Japanese-made items that made their way into Finland were toys, but beginning in the 70’s there is an increase in the import of high-tech products. Both visualizations are scrollable where they do not fit in the screen at once.

For particular domains, special presentations particularly suited to them are available [1]. In the current CultureSampo system, these are a timeline visualization for the time facet and a map visualization for the place facet. As an example of the use of the timeline visualization, Figure 3 depicts all items depicting beards and relating to Finland on a timeline, from which the user may discern if there was any change in beard styles in Finland near the end of the 19th century. Figure 4 on the other hand depicts a search for anything relating to churches on a map. This presentation can be used for example to infer information on the distribution of churches in southern Finland. In both of these visualizations, if a second dimension of organization is specified, it is expressed by marker coloring.

5.3. *Thematic views*

While the exhibition generation view presented before is very powerful, it can also be quite overwhelming for a first time user. To address this, the CultureSampo portal provides not only a single massively user-configurable view-based interface, but also a selection of expert pre-selected views to the data, based on thematic viewpoints [24].

Apart from one, these views provide pre-selected and pre-configured subsets of the complete search and organize functionality, along with small bits of additional tuned functionality. For example, in the history view of CultureSampo, a preselected query returns only historical events, and the user is left with choosing from pre-configured timeline and list view visualizations. Because these thematic views are based mostly on common general functionality, it is easy to add more of them based solely on the recommendations of content access specialists.

In this way, these views are very closely comparable to traditional physical exhibitions, with items and information pre-selected and pre-organized into various forms of thematically interesting and informative displays by cultural heritage institution curators. The idea here is that conceptual work of selecting interesting materials and viewpoints to that material can be trans-

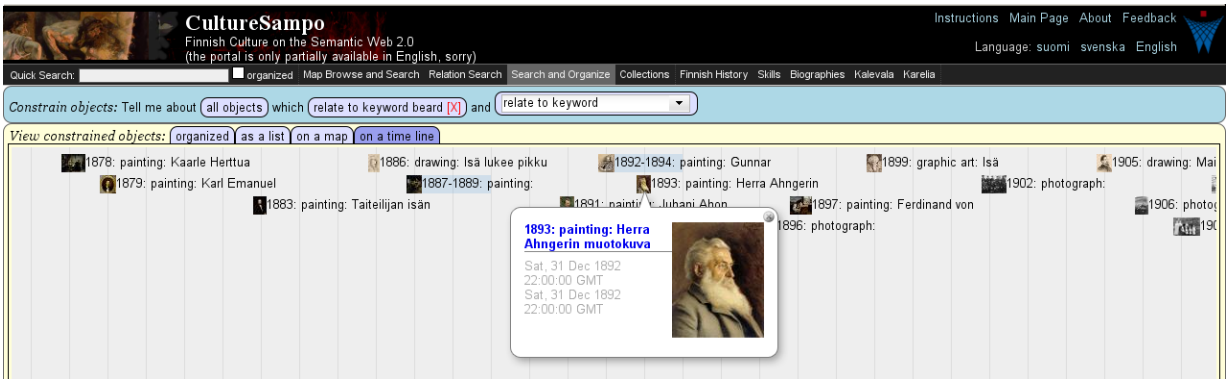

Fig. 3. Timeline visualization in CultureSampo

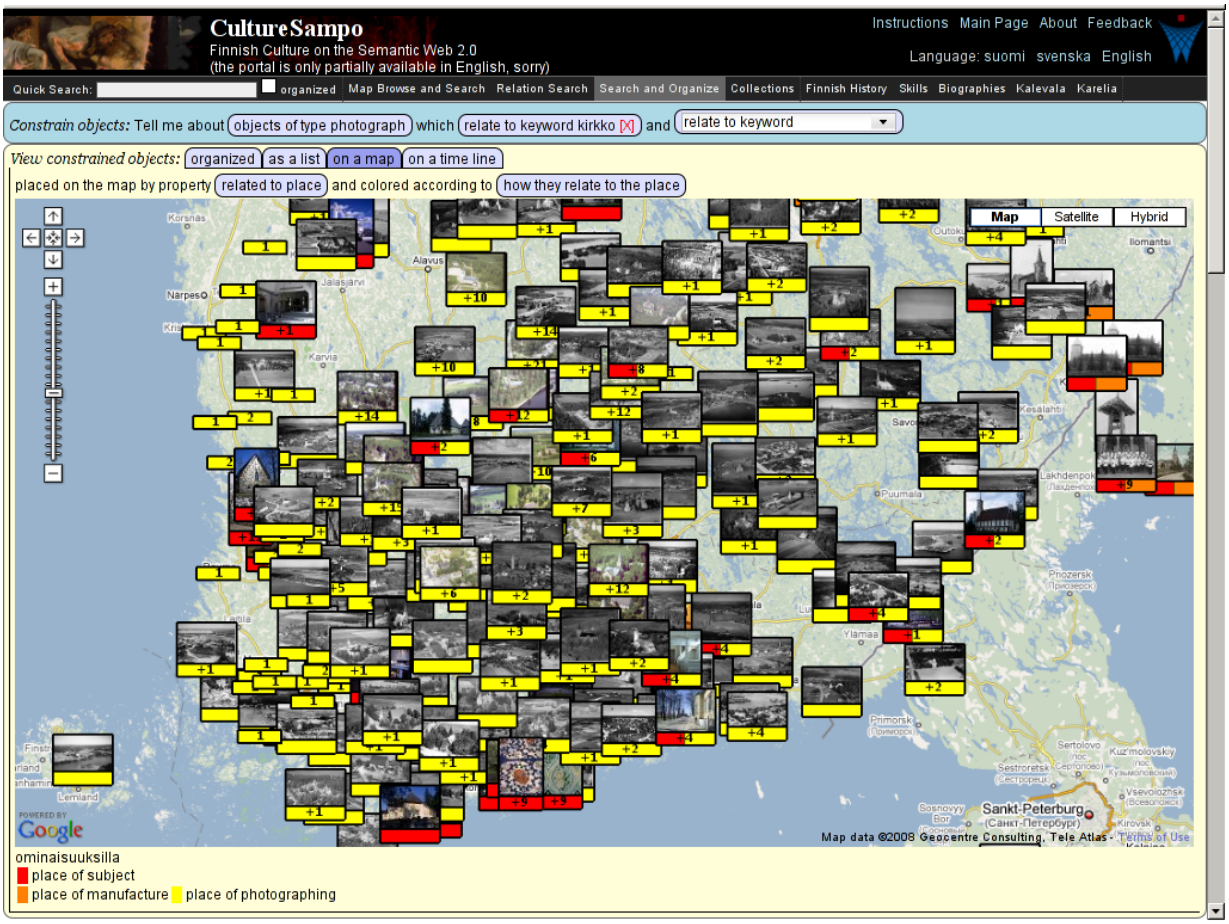

Fig. 4. Map visualization in CultureSampo

ferred from the user into the hand of experts, thereby giving easier points of access to the content.

While described in detail in a separate prior publication [24], a few of these views will be highlighted here in order to elucidate the concept.

5.3.1. Historical areas and maps

The historical areas and maps views of CultureSampo can be used for finding historical Finnish counties and places on modern maps, while at the same time linking to other cultural content related to those

places. The historical areas themselves are visualized using transparent historical maps on top of contemporary Google Maps, as in Figure 5, or as polygons capable of showing changes in the borders of the historical area through time, as in Figure 6 [36]. These latter are available through work done on the SAPO spatiotemporal Finnish place ontology [36,29]. Other view components included are a simple selection list listing all historical areas, shown on the left, and an organized list view listing all items related to the currently selected historical place organized by their semantic relation to that place.

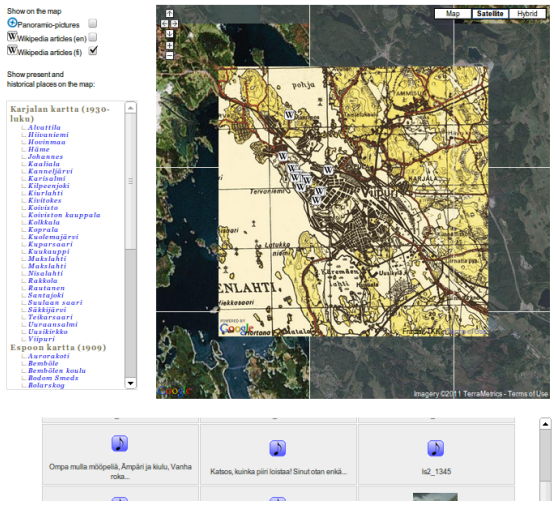

Fig. 5. Historical maps as viewed in CultureSampo

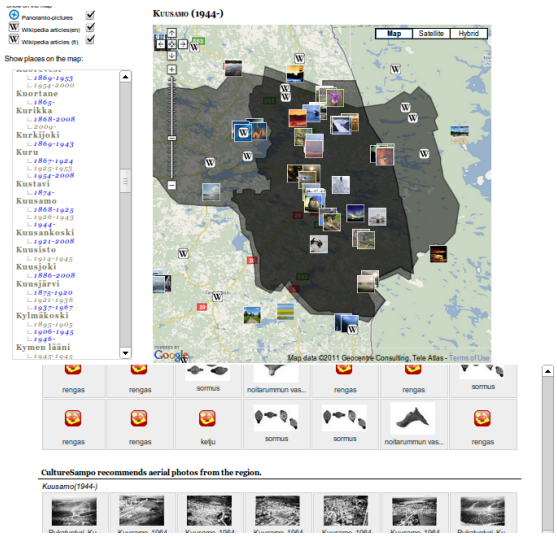

Fig. 6. Historical borders as viewed in CultureSampo

5.3.2. *Culture nearby*

In the culture nearby view, the geographical location of the user is utilized for finding cultural sites nearby, which the user might want to visit. Shown are nearby museums, architecturally and historically important sites as well as nature sites, organized in a list by distance to the user, as well as on a map.

5.3.3. *Connection discovery*

There is one view which cannot be described as offering a subset of the search and organize functionality, and is thus interesting when discussing also the limitations of the view-based search approach. This is the relational search view [38]. Here, the user can enter the names of information resources, and is returned with a description of how they are related to each other.

The key difference between this and the other views is that here the interesting information items are the paths between content items and not a set of those content items themselves organized in some way.

In CultureSampo, a subset of the Getty ULAN registry of 120,000 artists and organizations with 390,000 names was used. Here the user can type in two names, using semantic autocompletion, and CultureSampo tells how the persons or organizations are related to each other by the social network based on some 50 different social roles (e.g., parent-of, teacher-of, patron-of etc.). The underlying network can also be browsed by a graphical network browser. For example, in Figure 7 the user has typed in Napolean I (the French emperor) and Akseli Gallen-Kallela (a Finnish painter), and CultureSampo has found a social path of 7 steps between the persons. The browseable social network of Napolean I is depicted on the right hand side window.

While this view currently only shows the relations between single individuals, we have also later had success with displaying aggregate relations based on search patterns, such as for example visualizing the flight of European artists in the 20th century to the United States on a map, or similarly visualizing the import and export patterns of different types of items [35].

5.3.4. *Other views*

In addition to the views discussed, the portal contains also a view organizing the content of the portal by collection authority, in essence creating an automatically generated home page for each organization in the system with links to its collections and the collection items.

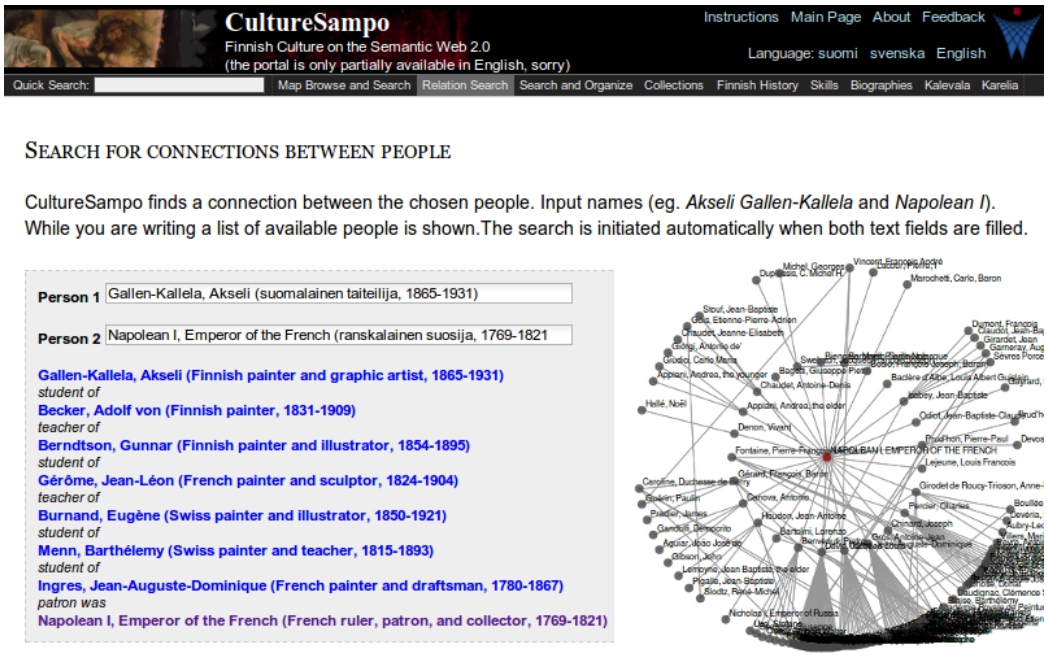

Fig. 7. Answering the question of how Napolean I, the French emperor is related to Akseli Gallen-Kallela by relational search in CultureSampo

Other views are the history timeline view already mentioned, as well as views that highlight and organize certain key content items or item types in the portal. These are the Kalevala view, which allows viewing a semantically annotated version of the epic, a view listing key biographies stored in the portal, the Karelia view, which lists automatically semantically enriched Finnish Wikipedia articles on Karelia which have been linked to other content in CultureSampo, and a view that highlights examples of skills and cultural narratives as content that can be semantically described and placed under search. [24]

6. Exposing functionality as web services

According to the vision of CultureSampo, a semantic cultural heritage portal should not only harvest and aggregate content for itself, but also provide it back to its content providers semantically enriched. What is needed is a web service that enables external (machine) users to take the combined content and functionality back, and embed customized versions of it inside the web pages of the participating organizations. In this use scenario, CultureSampo is utilized like Google Maps in mashup applications. We also wanted to support external users in building completely new custom user interfaces on top of the core CultureSampo services. In order to support this, several different web service functionalities were exposed. These correspond to standard interfaces common to many linked open data and semantic web applications. However, for some of the functionalities needed, no standards were yet available, so custom interfaces needed to be created.

6.1. RDF browsing service

As a first interface, the portal supports querying for the RDF related to a particular resource by a URL. This interface can be used by semantic web browsers such as Tabulator[21] or Disco[22] to browse the content of the portal, or as a web service interface for getting all the information necessary to display a particular item in a remote user interface.

Because the content published in CultureSampo comes from many different sources and is thus in different namespaces not under the CultureSampo domain, special support is needed to support the semantic web browsers in utilizing the portal. Here, for every resource URI in the RDF response, an additional rdfs:seeAlso link is generated in the returned RDF,

[21] http://www.w3.org/2005/ajar/tab
[22] http://sites.wiwiss.fu-berlin.de/suhl/bizer/ng4j/disco/

pointing to the CultureSampo description service URL for that resource.

In order to support the use of the web service for building remote user interfaces, the web service doesn't return just the symmetric concise bounded description [54] of the information item queried, but also includes all type and label statements of any resources that appear in that description, so those can be shown in the user interface in place of the URIs. To allow for even further customization, the systems can be configured in the data using a custom RDF path language to return even further resources. This additional support was installed based upon a need encountered in the BookSampo system, discussed later in this paper.

6.2. SPARQL service

As a general query interface, the portal supports SPARQL. All SPARQL queries are ran against the inferred unified database so that those doing the querying need not deal with equivalency or subsumption handling. However, only original triples and resources will be returned as variable bindings, so the client is not swamped with automatically inferred results. In CultureSampo's own recommendation rules written in SPARQL, this allowed vast simplification of the rules without sacrificing either recall or precision. For example, a prior version of the rules had to enumerate all the sub-properties of "foaf:knows", as well as all sub-types of "foaf:Agent" when doing social network recommendation. Hiding all this under the query layer made writing rules much simpler.

This added recall is also available for outside SPARQL clients, so that for example pointing the RelFinder application [15] to a DBPedia [6] instance loaded into the CultureSampo store finds more relations between entities than when using the standard DBPedia SPARQL endpoint.

The SPARQL service of CultureSampo is used also in creating customized JavaScript recommendation widgets to be embedded in the web pages of the content providers. For example, Figure 8 demonstrates how the functionality is used to bring example works by the painter Akseli Gallen-Kallela to be shown next to his biography at the Finnish Literature Society.

6.3. View-based and text search services

Because SPARQL isn't really suited for efficient text or view-based search, the portal provides a custom web service to support these tasks. The interface is a simplification of the one used in the Ontogator service [42] which was the view-based search service used in the precursor to CultureSampo, MuseumFinland [26].

For efficiently returning all information required to show the search results in a user interface, the service supports the same RDF path language definitions for returned content as the RDF browsing interface.

6.4. Geo-search service

For use primarily in various mobile use cases, the CultureSampo engine supports querying by geolocation, either using a bounding box or by a point and distance. This functionality has been used for example in the SmartMuseum EU project [49] to bring cultural recommendations to a PDA device, as well as in the Semantic Ubiquitous Services project[23] aiming at creating personalized, mobile cultural heritage services for mobile phones, leveraging material from for example DBPedia, OpenStreetMap and CultureSampo.

The geo-search service has also been hooked to the Layar API[24], which allows the content to be shown in the Layar augmented reality browser available on iPhones and Android devices. Another export format for the data is KML[25], for use in Google Earth and GPS nagivators.

6.5. Recommendation service

Finally, the CultureSampo service also includes a dedicated recommendation engine. This engine [49], which was developed in the SmartMuseum EU project operates on a profile of RDF triples, and tries to find clusters of other content related in interesting ways to that profile. The profile that is fed into the engine can be either the annotations of a content item, which can be used by content providers to find related content to that of their own, or alternatively an RDF-based user profile, with triples exposing preferences of the user. The latter form was used in the SmartMuseum project [49] to automatically create custom guided museum and culture site tours based on user profiles.

[23] http://www.seco.tkk.fi/projects/subi/
[24] http://site.layar.com/create/
[25] http://code.google.com/apis/kml/documentation/

Fig. 8. Content recommendations from CultureSampo in the far right column of the interface have been brought to the page using SPARQL and JavaScript

7. Use case: BookSampo

To concretize the different aspects of CultureSampo presented before, we will now walk through a concrete project which utilizes particularly many of the functionalities of CultureSampo to support the creation of a completely separate site. The project, BookSampo, is an ongoing joint venture of the Finnish public libraries to create a centralized web portal for all things related to fiction literature published in Finland in either Finnish or Swedish.

The co-operation between CultureSampo and BookSampo began when the libraries approached the CultureSampo project group. The projects seemed a good fit for each other, as literature was a good addition to the content of CultureSampo, and on the other hand many places were identified where the CultureSampo tools could be used for the good of BookSampo.

First, the BookSampo project itself needed to do data integration. Data on books was to be sourced primarily from the HelMet cataloguing system[26] used in the Helsinki metropolitan area libraries, but that system only contained subject keywords for fiction published since the year 1997, while some collections in smaller Finnish libraries could be used for getting keywords to older books, too. Also, from very early on, the vision of the project included storing information not only on the books, but also on the authors of those books. Data on these were to be sourced from three different databases maintained at various county libraries across Finland. Thus, the project already had at least two quite different content types, and multiple data sources.

As BookSampo didn't have a data store or editor environment of their own ready, the project decided to store their data natively as RDF, and adopt the SAHA RDF-based metadata editor [39] developed by the FinnONTO project as their primary editing environment.

Converters for the different data sources were then created, and the results mapped using the CultureSampo tools to each other as well as to other sources in CultureSampo, such as places, people and organizations. To link the keywords describing the literature to other content in CultureSampo, linked Finnish and Swedish thesauri for fiction indexing (Kaunokki and Bella) were converted into a bilingual ontology and linked with the upper ontology YSO [50].

This already bought instant benefit to the project, as 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

[26] http://www.helmet.fi/search S9/

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.

The BookSampo project also benefited from the close integration between SAHA and the ONKI ontology services [55] developed in the FinnONTO project. Fields in the SAHA editor were linked with semantic autocompletion widgets [23] that allowed selection from choice ontologies and instance registries in the KOKO infrastructure. This allowed also the ongoing manual indexing of further material to make use of the place and author registries, as well as increased the quality and sped up the look-up of content description keywords in the fiction description ontology.

Recently, the 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[27] used also for HelMet, automatically converted to RDF using the CultureSampo tools, and added to the SAHA project with tags indicating they should be verified. The librarians then use the SAHA support for this task, removing the "unverified" tags as they go along.

The flexibility of the RDF data model as well as the editor has also proved an asset. 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 when the whole CultureSampo model was likewise altered. 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, this model of a single object worked well, until it was decided that the project should also extend to include Swedish language literature. This necessitated a rethinking of the simple model [19] to take on more levels from the FRBRoo model [46] for bibliography information. As a result, works are currently described at two levels in BookSampo: 1) as an abstract work, which refers to the contents of the work that is the same in all translations and versions, and 2) as a physical work, which describes features inherent to each translation or version. Due to the flexibility of the RDF data model and editor, the transformation to this model could be done by running quite a simple programmatic transformation.

This complication of the model however 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[28] portal system. However, the intention of the project has always been that the search and recommendation functionality, as well as all primary information is kept and served solely 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 not efficiently deal with items where core content was split into multiple RDF resources. 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 to render a complete visualization of the book in a user interface. For example, the covers of the books are modeled as separate resources so that keywords and designer information can be associated with them. Because covers provide a good discerning visual element to be shown next to e.g. search results, there needed to be a way to efficiently include these objects in the material the CultureSampo search interfaces returned.

On the item page of the BookSampo portal, even more information is required, such as information on the publication series the book belongs to, any prizes the book has received, information on the author and the publisher, as well as core information on books that have been indexed by librarians as recommended for the readers of the currently viewed book. Now, in the data model of BookSampo, these are all stored as distinct objects.

Because of network latency issues, it is important to be able to get all this required information in one

[27] http://www.loc.gov/standards/marcxml/

[28] http://drupal.org/

query round-trip. Originally, it was thought that the item page could easily be constructed with SPARQL CONSTRUCT queries. However, when construction of such queries was actually attempted, it soon became apparent that the queries quickly grew too complex. For example, only getting all basic book details from the abstract and concrete work object along with the cover image resulted in a SPARQL query with a total of five OPTIONAL clauses, three of them nested. This was because it could not be assumed that all abstract books had covers, or even physical edition information.

To solve this, the project moved from using the SPARQL interface to use the general RDF browser interface, which was extended with support for specifying the resources and properties to return for a book or author URI as RDF path expressions. Currently, the end-user interface of the BookSampo project is at the phase were individual item pages function, and work is now moving on applying and testing the search functionality of CultureSampo in the portal.

8. Conclusions

In this paper, the lessons learned in developing the CultureSampo system for integrating and publishing heterogeneous linked data were presented. In the following, these will be summarized for convenience.

First, for integrating heterogeneous data, two frameworks were created in the project. Of these, the event-based primitive framework first attempted was found to function adequately for inference and recommendation, but to perform poorly in user interfaces and understandability. The later attempted more traditional ontology and schema mapping approach proved sufficient when combined with advances in user interface development. In addition, it allowed more ready leveraging of existing ontology and schema mapping tools. To attain a high quality of integration, an important part of this work was the large scale mapping of the domain ontologies used to each other on a national scale.

As regards converting and publishing legacy cultural heritage data as linked data, in addition to the needs for mapping the vocabularies and reference databases used, problems were identified as mostly arising from poorly designed original database systems as well as the proliferation of text fields for recording vocabulary references. The four part pipeline architecture presented in this paper allows tackling each step in the conversion process separately, but attacking the problems in those steps maximally globally across different content types and sources formats.

Regarding the added value that can be gained from semantically linking heterogeneous collections, we have argued the need for a shift of focus in semantic search from item location to presentation generation and support for exploratory search. In particular, we argue that often what is interesting in semantic databases are not the items themselves, but how they shed light on a theme described by a particular combination of domain concepts.

For the cultural heritage domain, museum exhibitions offer a suitable parallel to this idea. The search and organize view of CultureSampo takes advantage of this, combining an intuitive, yet expressive exhibition generation interface with different kinds of exhibition visualizations. On the exhibition generation side, a major contribution is the narrative query pattern for forming exhibitions combined with the concept of domain-centric view-based search, which allow catering to both searching for items having particular properties, as well as pure domain exploration.

On the exhibition visualization side, a simple, general-purpose visualization was created, as well as complemented with special purpose visualizations. Even these simple visualizations already give significant support for a user wanting to make sense of the data.

However, here there is still also much more that could be done. Our next user interface functionality will be to allow the user to select some rows or columns from the matrix for specific comparison and study. It may also be possible to aid such comparison work by automatically extracting from the data meaningful differences and similarities between neighboring exhibition rooms, such as “18th century agricultural items are more often made of wood than 19th century ones”.

In addition to allowing the users to create their own exhibitions, the CultureSampo portal also provides expert pre-configured views to the data. These allow visually highlighting particularly interesting topics in the content, as well as provide an easy entry point into the data for casual users.

A functional requirement that emerged from the CultureSampo project was that the portal should not just assimilate content from different organization, but should provide its advanced functionality back to them, so that the participating organizations could create mash-ups and custom interfaces to the data on their own. Here, it was discovered that the current standard

interfaces for linked data browsing and querying were not yet powerful enough for many needs, thus necessitating the creation of custom interfaces to enable those functionalities.

Finally, the use case of BookSampo was used to concretize the work done, as well as highlight the benefits an outside organization can now gain from participating in the CultureSampo infrastructure.

Acknowledgements

This work is part of the National Semantic Web Ontology project in Finland FinnONTO[29] (2003-2012), funded mainly by the National Technology and Innovation Agency (Tekes) and a consortium of 38 public organizations and companies.

The authors wish to thank all the researchers, cultural institution staff and everyone else involved in the various iterations of the CultureSampo and MuseumFinland projects. Specifically, thanks must go to Tomi Kauppinen, Olli Alm, Jussi Kurki, Joeli Takala, Kimmo Puputti, Heini Kuittinen, Tuomas Palonen, Panu Paakkarinen, Joonas Laitio, Katariina Nyberg, Jari Väätäinen and Jouni Hyvönen, who all took part in the creation of the final version of the CultureSampo portal.

References

[1] Omar Alonso, Ricardo Baeza-Yates, and Michael Gertz. Exploratory search using timelines. In *Proceedings of the SIGCHI 2007 Exploratory Search and HCI workshop*, 2007.

[2] Lorin W. Anderson and David A. Krathwohl, editors. *A taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Allyn & Bacon, Boston, Massachusetss, 2000.

[3] Nicholas J. Belkin, Pier Giorgio Marchetti, and Colleen Cool. Braque: design of an interface to support user interaction in information retrieval. *Informationg Processing and Management*, 29(3):325–344, 1993.

[4] S. Bergamaschi, S. Castano, and M. Vincini. Semantic integration of semistructured and structured data sources. *SIGMOD Rec.*, 28:54–59, March 1999.

[5] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data – the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2009.

[6] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia - a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154 – 165, 2009.

[7] Namyoun Choi, Il-Yeol Song, and Hyoil Han. A survey on ontology mapping. *SIGMOD Rec.*, 35:34–41, September 2006.

[8] Chun Wei Choo, Brian Detlor, and Don Turnbull. Information seeking on the web: An integrated model of browsing and searching. *First Monday*, 5(2), February 2000.

[9] P. F. Cole. The analysis of reference query records as a guide to the information requirements of scientists. *Journal of Documentation*, 14(4):197–207, 1958.

[10] Colleen Cool and Nicholas J. Belkin. A classification of interactions with information. In Harry Bruce, Ray Fidel, Peter Ingwersen, and Pertti Vakkari, editors, *Emerging frameworks and methods; Proceedings of the 4th international conference on conceptions of Library and Information Science (COLIS4)*, pages 1–15, Greenwood Village, CO, July 2002. Libraries Unlimited.

[11] Isabel F. Cruz, Stefan Decker, Dean Allemang, Chris Preist, Daniel Schwabe, Peter Mika, Michael Uschold, and Lora Aroyo, editors. *The Semantic Web - ISWC 2006, 5th International Semantic Web Conference, ISWC 2006, Athens, GA, USA, November 5-9, 2006, Proceedings*, volume 4273 of *Lecture Notes in Computer Science*. Springer, 2006.

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

[13] Marti A. Hearst, Jennifer English, Rashmi Sinha, Kirsten Swearingen, and Ping Yee. Finding the flow in web site search. *Communications of the ACM*, 45(9):42–49, September 2002.

[14] Tom Heath and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space (1st edition)*. Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool.

[15] Philipp Heim, Steffen Lohmann, and Timo Stegemann. Interactive relationship discovery via the semantic web. In Lora Aroyo, Grigoris Antoniou, Eero Hyvönen, Annette ten Teije, Heiner Stuckenschmidt, Liliana Cabral, and Tania Tudorache, editors, *The Semantic Web: Research and Applications*, volume 6088 of *Lecture Notes in Computer Science*, pages 303–317. Springer Berlin / Heidelberg, 2010.

[16] Michiel Hildebrand, Jacco van Ossenbruggen, and Lynda Hardman. /facet: A browser for heterogeneous semantic web repositories. In Cruz et al. [11], pages 272–285.

[17] Michiel Hildebrand, Jacco van Ossenbruggen, and Lynda Hardman. An analysis of search-based user interaction on the semantic web. Technical report, Centrum voor Wiskunde en Informatica (NL), 2007.

[18] David F. Huynh, David R. Karger, and Robert C. Miller. Exhibit: lightweight structured data publishing. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 737–746, New York, NY, USA, 2007. ACM.

[19] Kaisa Hypén and Eetu Mäkelä. An ideal model for an information system for fiction and its application: Kirjasampo and semantic web. *Library Review*, 60(4), April 2011. Forthcoming.

[20] Eero Hyvönen. Semantic portals for cultural heritage. In Steffen Staab and Rudi Studer, editors, *Handbook on Ontologies (2nd Edition)*. Springer-Verlag, 2009.

[21] E. Hyvönen, M. Junnila, S. Kettula, E. Mäkelä, S. Saarela, M. Salminen, A. Syreeni, A. Valo, and K. Viljanen. Finnish Museums on the Semantic Web. User's perspective on MuseumFinland. In *Proceedings of Museums and the Web 2004 (MW2004), Seleted Papers, Arlington, Virginia, USA*, 2004.

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

http://www.archimuse.com/mw2004/papers/hyvonen/hyvonen.html.

[22] 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.

[23] Eero Hyvönen and Eetu Mäkelä. Semantic autocompletion. In *Proceedings of the first Asia Semantic Web Conference (ASWC 2006), Beijing*. Springer-Verlag, New York, August 4-9 2006.

[24] Eero Hyvönen, Eetu Mäkelä, Tomi Kauppinen, Olli Alm, Jussi Kurki, Tuukka Ruotsalo, Katri Seppälä, Joeli Takala, Kimmo Puputti, Heini Kuittinen, Kim Viljanen, Jouni Tuominen, Tuomas Palonen, Matias Frosterus, Reetta Sinkkilä, Panu Paakkarinen, Joonas Laitio, and Katariina Nyberg. CultureSampo – Finnish culture on the semantic web 2.0. Thematic perspectives for the end-user. In *Proceedings, Museums and the Web 2009, Indianapolis, USA*, April 15-18 2009.

[25] Eero Hyvönen, Eetu Mäkelä, Tomi Kauppinen, Olli Alm, Jussi Kurki, Tuukka Ruotsalo, Katri Seppälä, Joeli Takala, Kimmo Puputti, Heini Kuittinen, Kim Viljanen, Jouni Tuominen, Tuomas Palonen, Matias Frosterus, Reetta Sinkkilä, Panu Paakkarinen, Joonas Laitio, and Katariina Nyberg. CultureSampo – a national publication system of cultural heritage on the semantic web 2.0. In *Proceedings of the 6th European Semantic Web Conference (ESWC2009), Heraklion, Greece*, 2009. Springer-Verlag.

[26] Eero Hyvönen, Eetu Mäkelä, Mirva Salminen, Arttu Valo, Kim Viljanen, Samppa Saarela, Miikka Junnila, and Suvi Kettula. MuseumFinland – Finnish museums on the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 3(2-3):224 – 241, 2005. Selcted Papers from the International Semantic Web Conference, 2004 - ISWC, 2004.

[27] Eero Hyvönen, Eetu Mäkelä, Mirva Salminen, Arttu Valo, Kim Viljanen, Samppa Saarela, Miikka Junnila, and Suvi Kettula. 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.

[28] Eero Hyvönen, Tuukka Ruotsalo, Thomas Häggström, Mirva Salminen, Miikka Junnila, Mikko Virkkilä, Mikko Haaramo, Eetu Mäkelä, Tomi Kauppinen, and Kim Viljanen. CultureSampo – Finnish culture on the semantic web: The vision and first results. In Klaus Robering, editor, *Information Technology for the Virtual Museum – Museology and the Semantic Web*, pages 33–58. LIT Verlag, Berlin, November 2007.

[29] Eero Hyvönen, Jouni Tuominen, Tomi Kauppinen, and Jari Väätäinen. Representing and utilizing changing historical places as an ontology time series. In Naveen Ashish and Amit Sheth, editors, *Geospatial Semantics and Semantic Web: Foundations, Algorithms, and Applications*. Springer-Verlag, 2011, forth-coming.

[30] Thomas Häggström. Toimintakeskeisen semanttisen monināykymähaun toteutus ja evaluointi kulttuurialan portaalisovelluksessa. Master's thesis, Helsinki University of Technology (TKK), December 2007.

[31] Bernard J. Jansen, Brian Smith, and Danielle Booth. Learning as a paradigm for understanding exploratory search. In *Proceedings of the SIGCHI 2007 Exploratory Search and HCI workshop*, 2007.

[32] M. Junnila, E. Hyvönen, and M. Salminen. Describing and linking cultural semantic content by using situations and actions. In Klaus Robering, editor, *Information Technology for the Virtual Museum*. LIT Verlag, 2007.

[33] Yannis Kalfoglou and Marco Schorlemmer. Ontology mapping: the state of the art. *Knowl. Eng. Rev.*, 18:1–31, January 2003.

[34] Tomi Kauppinen, Glauco Mantegari, Panu Paakkarinen, Heini Kuittinen, Eero Hyvönen, and Stefania Bandini. Determining relevance of imprecise temporal intervals for cultural heritage information retrieval. *International Journal of Human-Computer Studies*, 86(9):549–560, September 2010.

[35] Tomi Kauppinen, Kimmo Puputti, Panu Paakkarinen, Heini Kuittinen, Jari Väätäinen, and Eero Hyvönen. Learning and visualizing cultural heritage connections between places on the semantic web. In *Proceedings of the Workshop on Inductive Reasoning and Machine Learning on the Semantic Web (IRMLeS2009), The 6th Annual European Semantic Web Conference (ESWC2009)*, May 31 - June 4 2009.

[36] Tomi Kauppinen, Jari Väätäinen, and Eero Hyvönen. Creating and using geospatial ontology time series in a semantic cultural heritage portal. In *S. Bechhofer et al.(Eds.): Proceedings of the 5th European Semantic Web Conference 2008 ESWC 2008, LNCS 5021, Tenerife, Spain*, pages 110–123. Springer-Verlag, June 1-5 2008.

[37] Phokion G. Kolaitis. Schema mappings, data exchange, and metadata management. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '05, pages 61–75, New York, NY, USA, 2005. ACM.

[38] Jussi Kurki and Eero Hyvönen. Relational semantic search: Searching social paths on the semantic web. In *Poster Proceedings of the International Semantic Web Conference (ISWC 2007), Busan, Korea*, Nov 2007.

[39] Jussi Kurki and Eero Hyvönen. Collaborative metadata editor integrated with ontology services and faceted portals. In *Workshop on Ontology Repositories and Editors for the Semantic Web (ORES 2010), the Extended Semantic Web Conference ESWC 2010, Heraklion, Greece*. CEUR Workshop Proceedings, http://ceur-ws.org/, June 2010.

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

[41] Eetu Mäkelä. *View-Based User Interfaces for the Semantic Web*. PhD thesis, Aalto University, School of Science and Technology, Espoo, November 2010. D.Sc. dissertation.

[42] Eetu Mäkelä, Eero Hyvönen, and Samppa Saarela. Ontogator - a semantic view-based search engine service for web applications. In Cruz et al. [11], pages 847–860.

[43] Eetu Mäkelä, Kim Viljanen, Olli Alm, Jouni Tuominen, Onni Valkeapää, Tomi Kauppinen, Jussi Kurki, Reetta Sinkkilä, Teppo Kansälä, Robin Lindroos, Osma Suominen, Tuukka Ruotsalo, and Eero Hyvönen. Enabling the semantic web with ready-to-use web widgets. In Lyndon J. B. Nixon, Roberta Cuel, and Claudio Bergamini, editors, *Proceedings of the Workshop on First Industrial Results of Semantic Technologies, co-located with ISWC 2007 + ASWC 2007, Busan, Korea, November 11th, 2007*, volume 293 of *CEUR Workshop Proceedings*, pages 56–69. CEUR-WS.org, 2007.

[44] Eyal Oren, Renaud Delbru, and Stefan Decker. Extending faceted navigation for RDF data. In Cruz et al. [11], pages 559–572.

[45] Antonella Poggi, Domenico Lembo, Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, and Riccardo Rosati. Linking data to ontologies. *J. Data Semantics*, 10:133–173, 2008.

[46] 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.

[47] Tuukka Ruotsalo and Eero Hyvönen. An event-based approach for semantic metadata interoperability. In Karl Aberer, Key-Sun Choi, Natasha Fridman Noy, Dean Allemang, Kyung-Il Lee, Lyndon J. B. Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*, volume 4825 of *Lecture Notes in Computer Science*, pages 409–422. Springer, 2007.

[48] Tuukka Ruotsalo and Eero Hyvönen. A method for determining ontology-based semantic relevance. In *Proceedings of the International Conference on Database and Expert Systems Applications DEXA 2007, Regensburg, Germany*. Springer, September 3-7 2007.

[49] Tuukka Ruotsalo, Eetu Mäkelä, Tomi Kauppinen, Eero Hyvönen, Krister Haav, Ville Rantala, Matias Frosterus, Nima Dokoohaki, and Mihhail Matskin. Smartmuseum: Personalized context-aware access to digital cultural heritage. In *Proceedings of the International Conferences on Digital Libraries and the Semantic Web 2009 (ICSD2009)*, Trento, Italy, September 2009. Trento, Italy.

[50] Jarmo Saarti and Kaisa Hypen. From thesaurus to ontology: the development of the kaunokki finnish fiction thesaurus. *The Indexer*, 28:50–58(9), June 2010.

[51] mc schraefel, Max Wilson, Alistair Russell, and Daniel A. Smith. mSpace: improving information access to multimedia domains with multimodal exploratory search. *Communications of the ACM*, 49(4):47–49, 2006.

[52] Guus Schreiber, Alia Amin, Mark van Assem, Viktor de Boer, Lynda Hardman, Michiel Hildebrand, Laura Hollink, Zhisheng Huang, Janneke van Kersen, Marco de Niet, Borys Omelayenko, Jacco van Ossenbruggen, Ronny Siebes, Jos Taekema, Jan Wielemaker, and Bob J. Wielinga. Multimedian e-culture demonstrator. In *The Semantic Web - Proceedings of the 5th International Semantic Web Conference2006*, pages 951–958, November 5-9 2006.

[53] John F. Sowa. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Course Technology, 1 edition, August 1999.

[54] Patrick Stickler. CBD – Concise Bounded Description. W3C submission, W3C, 30 sep 2004. http://www.w3.org/Submission/2004/SUBM-CBD-20040930/.

[55] Kim Viljanen, Jouni Tuominen, and Eero Hyvönen. Ontology libraries for production use: The finnish ontology library service ONKI. In Lora Aroyo, Paolo Traverso, Fabio Ciravegna, Philipp Cimiano, Tom Heath, Eero Hyvönen, Riichiro Mizoguchi, Eyal Oren, Marta Sabou, and Elena Paslaru Bontas Simperl, editors, *The Semantic Web: Research and Applications, 6th European Semantic Web Conference, ESWC 2009, Heraklion, Crete, Greece, May 31-June 4, 2009, Proceedings*, volume 5554 of *Lecture Notes in Computer Science*, pages 781–795. Springer, 2009.

[56] Ubbo Visser. *Intelligent information integration for the Semantic Web*. Springer-Verlag, 2004.

[57] Ryen W. White, Gheorghe Muresan, and Gary Marchionini. Evaluating advanced search interfaces using established information-seeking models. *Information Processing and Management*, 2007.

[58] Tom D. Wilson. Current awareness services and their value in local government. *Journal of Librarianship and Information Science*, 14(4):279–288, 1982.

[59] Tom D. Wilson. Information needs and uses: fifty years of progress. In Brian Campbell Vickery, editor, *Fifty years of information progress: a Journal of Documentation review*, pages 15–51, London, 1994. Aslib.