

Discovering Places of Interest through Direct and Indirect Associations in Heterogeneous Sources — The TravelSampo System

Eetu Mäkelä, Aleksi Lindblad, Jari Väättäinen, Rami Alatalo, Osma Suominen, and Eero Hyvönen

Semantic Computing Research Group (SeCo),
Aalto University and University of Helsinki, Finland
first.last@aalto.fi, <http://www.seco.tkk.fi/>

Abstract. Linked data related to places offers a superior collection of information to base search and recommendation functionality on in eTourism visit planning as well as location-aware mobile applications. Besides places interesting in themselves, through linked data it is possible to discover places interesting only through association, such as being the venue for a concert by an artist with an interesting genre. However, in order to harness this collective data source, challenges relating to data heterogeneity, quality, scale, and indexing and querying complexity must be resolved. In this paper, the TravelSampo visit planning and mobile application is presented, which tackles these issues. Using the system, queries describing both simple and complex interests can be run over some 17 million places of interest from over 20 vastly heterogeneous sources.

1 Introduction

Location-aware mobile devices are becoming increasingly commonplace. This has led to a multitude of mobile applications to search for e.g. events, places of interest or services near the user's physical location. On the other hand, many eTourism web applications also now allow people to design travel plans online, picking sites to visit and exporting visit lists to their phone's navigator software.

The TravelSampo project is an attempt to harness linked data as a source of material for an application to help travellers find content relevant to them, both in planning as well as during a trip. As compared to existing non-linked data solutions as well as similar linked data systems such as DPBedia Mobile [2], mSpace Mobile [13] and SmartMuseum [11], it tries to improve upon the state of the art in being able to integrate both massively more heterogeneous material, as well as to provide more intelligent services on top of it.

Particularly, the TravelSampo system takes into account that there are multiple ways in which a location may be of interest to a user. First, the place itself may have some quality of interest, such as being a church, or being a church in the gothic style. On the other hand, the place may also be of interest only

through a more or less direct association, such as being the venue for an interesting event or having been the birthplace of a painter with a style of interest. In addition, a place may be of interest by virtue of the services offered there, such as Internet access.

This variety of ways in which data can be both interesting as well as encoded necessitates a flexible architecture for querying locations of interest. The strength of this is that the application should ultimately be able to cater to a wide variety of interests, from people looking for nearby museums through music fans interested in concerts by Norwegian heavy metal bands to freegans searching for dumpsters near big supermarkets without nearby surveillance cameras. At the same time, this scale of heterogeneity causes severe problems in both integrating the content as well as providing efficient and intelligent search and recommendation services and user interfaces on top of it.

In the current demonstration system of TravelSampo, some 17 million locations have been loaded into the system, integrating information from over 20 vastly different datasets of places, places of interest, and content making places interesting through association, such as fiction taking place in real-world locations, or the birthplaces of famous artists. Included are for example the huge datasets of DBPedia [3] and the LinkedGeoData.org [1] version of OpenStreetMap, but also fast-changing, dynamically converted datasets such as four different sources for current events and exhibitions in Finland.

In this paper, the TravelSampo application is presented first through its user interface. After that, the challenges faced and solutions developed in integrating, mapping and making usable the disparate heterogeneous data sources are described. Finally described are the indexing and querying interfaces created that make possible the complex queries required to provide the advanced functionality of the TravelSampo application.

2 The TravelSampo Application

The TravelSampo application has two distinct interfaces. The web interface is used to plan the trip beforehand and to examine and share the trip afterwards. The mobile interface is used during the trip to find the destinations and to get more information about them.

2.1 The Visit Planning Interface

A typical user would be someone who is going on a trip to a new city. Before the trip he can use the planning interface to find out what kind of cultural destinations and events the city offers during his trip. The destinations can be searched with different levels of complexity. In the simplest case our user is interested in churches, which are places themselves. Our user is also interested in wall climbing, which is a service located in a place. And finally he's interested in modern art, which is a topic of an exhibition held in a place. The application can handle all these searches.

The user has found a couple of churches, a sport center with wall climbing and an exhibition of modern art and now he can save them to his destination shelf which can be accessed during the trip in the mobile interface. The filters used to find these destinations can also be saved to be used on other trips either in the planning interface or the mobile interface. The visit planning interface is not yet implemented but the shelves and filters can be produced and used in the mobile interface.

2.2 The Mobile Interface

During his trip the user can use the mobile application to find the previously saved destinations. The starting page of the mobile interface, depicted in figure 1 has a list of nearby destinations (1.A) with their types, distance to them and interesting instances associated to them which can be filtered using the filter menu (1.B). This menu contains the users destination shelves that are relevant in his current location as well as his personal filters and some predefined ones.

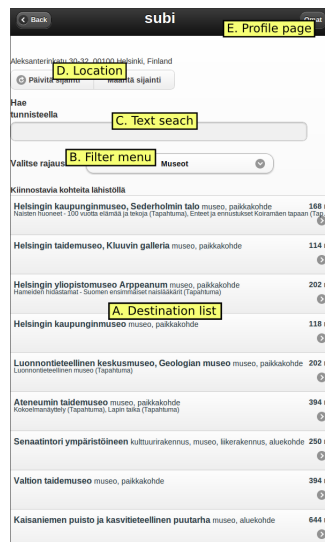


Fig. 1. Main screen of the mobile interface

The user can now use the shelf containing his destinations to access their pages and get a route map to the destination as well as information about it, links to associated instances and a button to mark the destination for future reference. As the user is in the destination page in the context of the destination shelf there are also buttons to browse through other destinations of the shelf.

If the user finds he has more time than he expected he can go back to the main page and use his saved filters to find for example more churches or in-

stances related to modern art. This can also be used to quickly find interesting destinations or useful services on the vicinity without prior planning.

There will also be a possibility to use free text search to find destinations (1.C). The location area (1.D) shows the user’s current address and allows him to reload it using the geolocation capabilities of the mobile device or set it manually. On the top right corner there is a link to the user’s profile page or a login page if the user hasn’t already logged in (1.E).

3 Data Sources and Modeling

As already stated, the TravelSampo data repository contains some 17 million locations sourced from over 20 vastly different datasets. These sources are described in table 1, stating the general type of data sourced from each provider, the number of location and reference items in the data sources, as well as some example content types in the data thus gained. As can be seen from the table, the TravelSampo system contains a truly heterogeneous mix of data of different types, sources and schemas.

Particularly interesting in analyzing the data sources integrated is the fact that the boundaries between geographic place names, places of interest and services are not crisp. For example, the general place name registries contain not only hills and swamps, but also areas of sporting services, churches and abandoned police posts, while on the other hand the locations in the Espoo nature site point of interest database are precisely hills and swamps of special interest. Further, the Helsinki City service database for example contains both office entities such as childcare services as well as located services such as swimming pools, which also feature in point of interest databases.

Based on this observation, the TravelSampo system was designed not to discern between these sources for locations at all, but to treat all location information as equal. This puts the burden of discovering whether a location is a point of interest to someone on the information gathered for that location. Among direct indicators of interest, primary among them is the type of the location. To be able to use this across all the place data in the TravelSampo system, it was decided to attempt to build a single unified place and place of interest type ontology POIO from all the type ontologies used in the various data sources.

This was done semiautomatically so that first all place type labels were compared automatically, already yielding several hundred equivalency mappings. Then, these mappings were examined by hand, and a large number of spurious mappings rejected, while an equally large amount of new mappings and subclass relations were curated, until all place types could be found under a single root. Table 2 relates the numbers of distinct place types in the constituents of the POIO ontology, as well as the size of the final ontology. In total, of the 2499 concepts in the final ontology, 62% (1539) were found to be shared between at least two source ontologies.

Besides differing in place types, the datasets also differed vastly in terms of modeling and level of content description. For example, the RKY database of

Table 1. TravelSampo Data Sources

Type	Source	Size	Example types	Description
Places of Interest	LinkedGeoData	~3.1 million	Statue, tunnel, crossing, school, ruins, bench	Entities with non-amenity types in the RDF conversion [1] of the open mapping project OpenStreetMap
	DBPedia	~432,000	Site of earthquake, tv-mast, bridge	RDF conversion [3] of structured information in Wikipedia
	MJREKI	25,343	Hill fort, rock tomb, place of village	Finnish ancient monuments
	RKY	1,850	Manor, powerplant, cemetery, rectory	Finnish nationally cultural-historically significant milieu
	Espoon rakenmukset	80	Villa, rectory, school, jugend building	Cultural-historically significant buildings in the city of Espoo
Services	Espoon luontokohteet	282	Swamp, glacial erratic, meadow, nature reserve	Nature sites in the city of Espoo
	LinkedGeoData	~3.7 million	Pub, pharmacy, restaurant	Entities tagged as amenities in the RDF conversion [1] of the global open mapping project OpenStreetMap
	Pääkaupunkiseudun palvelukartta	2,929←4350	Swimming pool, daycare, museum, child protection services	Public services provided by the cities in the Helsinki metropolitan area
	EvenemaX	~1480←4300	Pub quiz, concert, exhibit	Finnish commercial cultural event aggregator
	Turku 2011	~305←3400	Concert, exhibition, circus	Events included in Turku's year as European culture capital
Events	Museot.fi	~80←140	Exhibition	Current exhibitions at Finnish museums
	Agricola	~60←70	Lecture, seminar, exhibition	Event-calendar of the Agricola network of Finnish historians
	Tarinoiden Helsinki	1629←4091	Book, music, film, fact	Fiction and facts relating to places in Helsinki
Place-Related Content	CultureSampo	~20,000←600,000	Poem, photograph, painting, video, skill	An integrated portal of some 30 different content types from some 20 different institutions [6]
	DBPedia	~432,000←3.6 million	Person, organization, invention, event	Every conceivable notable aspect of human existence that linked to a place
	SUO	~800,000	Hill, swamp, meadow, glacial erratic, area of sporting services	Finnish registry of place names
Places	SAPO	1,261	Administrative area	A spatio-temporal ontology of historical Finnish counties [7]
	GNS	~4.2 million	Oil field, ramp, hill, abandoned police post, church, glacier	Geonet Names Server, a US government place database
	GeoNames	~6.9 million	Oil field, ramp, hill, abandoned police post, church, glacier	Open database of geographical names, using the same feature codes as GNS
Total	TGN	~895,000	Aqueduct, mausoleum, sinkhole, earl- dom, Nicaraguan center, dynasty	Getty thesaurus of historic and current locations
	Karelian places	37,476	Village, house	Historical places in the Karelia region of Finland and Russia
Total		~17 million←300 million		

Table 2. Number of distinct place types in the constituents of the POIO ontology

Name	Size
OpenStreetMap	1506
TGN	1737
GNS/Geonames	648
SUO	648+142 ^a
POIO	2499

^a GNS types+additions

culturally significant milieu contains areas of cultural interest defined as polygons. However, most of these areas are actually collections of multiple points of interest, which are not modeled separately at all. On the other hand, most of the other databases listed do not model areas at all, but only provide centerpoints for even large features. Even worse, it is often difficult to automatically deduce when a location actually refers to a notable mass of land, such as an amusement park, instead of a small point, such as a statue.

As regards services, in the vast majority of data and data sources used in the TravelSampo project, the services described are those that can be described indirectly through place type, such as being a restaurant or a pharmacy. However, in the Helsinki City data source the services offered at a particular location are described separately, for example noting if a particular library offers Internet access, has a scanner or loans AV equipment.

In the case of the TravelSampo system, particularly as it was making use of many automatically converted and dynamically updating data sources, it was decided that these heterogeneities in content modeling could not be unified, at least without losing information or the expressivity of the original data, but would have to be resolved at the query construction level. Fortunately, it seems that some quite general mapping rules could be made to facilitate this, for example linking services and events described as separate resources to the places they are provided or help in, or linking a culture site with no direct description to the compound description of the larger area it was found in.

As can be gleamed from the table listing the TravelSampo data sources, events such as concerts and exhibitions were identified as a particularly interesting non-direct element signaling a place of interest. That is, particularly for cultural applications, often one is for example not interested in a museum per se, but in the exhibitions that are on display in that museum at present.

Now, events are particularly dynamic sort of data. At the time of creating the TravelSampo system, there were no sources for current and coming event information in RDF. However, there were multiple sources from which such could be gleamed in other formats, such as comma separated values, JSON or RSS. The event content for TravelSampo thus comes from a converter pipeline that is capable of being run at regular intervals, or by request. This pipeline is actually

a more general one, called Harava [12], created in the FinnONTO project¹ [5] as a Semantic Web infrastructure tool by which data can be harvested, converted, enriched and validated to be published as quality Linked Data for anyone to use².

A major problem in the event data sources to be used in the project however was that none of them contained any machine-processable descriptions of the topics related to the event, such as the style of an exhibition or the artist. This problem was also evident in some of the point of interest data sources. For example, in the OpenStreetMap data on Helsinki, there is an object of type “memorial”, which only in its textual description says that 1) it is actually a statue and 2) it depicts the runner Paavo Nurmi.

To overcome these limitations, automated information extraction services were integrated into the TravelSampo architecture and the Harava pipeline, which could then extract relevant entities such as people, organizations and places as well as general content keywords from the textual descriptions of the events and other data items.

Because the information extraction tools were configured to use the whole vast TravelSampo database as a source for keywords, they are usually able to pick up a huge number of potentially related instances. The problem then became more of filtering these potential instances to the most important and factual ones. Fortunately, here the project could make use of the open source Maui information extraction tool [10], which has been previously shown to be human-competitive in selecting primary topic keywords from text.

4 Place and Event Instance Mapping

After getting all the different data sources together, one finds a large amount of overlap between them. Besides the same places occurring in many of the place databases, also events typically are entered in more than one of our dynamically updated event sources.

The indexing system used in the TravelSampo project is capable of inferring and resolving ontology language equivalency statements transparently. Thus, mapping between the different datasources does not need to happen at indexing time, but can be done centrally and iteratively in the global TravelSampo data space through generating RDFS, OWL and SKOS equivalencies. Actually, this task becomes 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 [8, 4] for doing just that.

Due to this, all but certain mappings are also relegated into this late stage of processing, so that for example all keywords found in the data sources during pipeline processing are created as resources in the data source’s own namespace, instead of being equated with ontology concepts directly. This also makes sure

¹ <http://www.seco.tkk.fi/projects/finnonto/>

² The source code of Harava has been released under a MIT style open source license and published as a Google Code project³.

that no information is lost and no errors introduced in indexing, due to e.g. the keywords used not being found in the reference vocabularies, or being translated to a wrong concept based on improper fuzzy reasoning.

As already said, the semantic enrichment done to the materials through information extraction tools 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 transparent resolution of equivalency statements also means that any erroneous mappings can be undone easily after the fact by just removing the RDF triple specifying the bad mapping. The tasks of verifying and improving the resource mappings generated, as well as verifying automatic enrichments, can be done in the TravelSampo ecosystem through the SAHA metadata editor [9] 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 validator service VERA⁴. 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.

In this way, the dynamically updated content of the TravelSampo portal can be iteratively improved and corrected as the system is running, right when problems are discovered, allowing focusing on the areas most critical to efficient use.

5 Indexing and Querying

Our stated choice of semantic integration by mapping properties and resources in RDF required that the triple-store used 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 TravelSampo, 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 subclasses 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.

⁴ <http://www.seco.tkk.fi/services/vera/>

In the material used for TravelSampo, 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 TravelSampo 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 any temporal entity resources such as event times and further resources related to them. These are all functionalities that were needed in the various user interfaces of the TravelSampo system. 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 TravelSampo, 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.

In order to cater to the complexity and heterogeneity of the data sources used in TravelSampo, the indexing and querying system also has to be able to efficiently query quite complex patterns. For example, the intent is that for example the query “Finnish electronica near Helsinki in the following week” would match a concert by Jimi Tenor at the Helsinki Ice Hall, because Jimi Tenor is a Finnish electronica artist playing there in the specified timeframe. However, inside the data model, this is quite a complex pattern, as visualized in figure 2.

To answer the query, first, each resource with a label matching any of the keywords must be found, as well as those matching the temporal and spatial constraints. This results in a result set with (among others) the nationality Finnish, the genre Electronica, the concert event and the location of Helsinki ice hall. Then, all resources relating to these or their subconcepts must be added to the result set. This results in (among others) the artist Jimi Tenor (who is Finnish) and the album Intervision (which has a genre of downtempo, which is a subgenre of electronica).

Finally, all resources that are not already locations must be mapped to any that they refer to, and finally an intersection taken between all locations found to reveal the final result. In this case, such mappings must also be done iteratively. While the concert event relates directly to the location, but the artist and the album are still two and three steps away, respectively. To obtain the final result, one must follow first the link from album to the artist, and then from the artist to the event, which then finally leads to the location.

To resolve this, the search functionality in the TravelSampo backend was split into multiple stages, each taking in SPARQL queries. First, multiple “select” queries are run, one for each incoming keyword, temporal and spatial constraint, acting on a dedicated index. Using the index, it is easy to efficiently return not only resources matching the spatial and temporal constraints, but also any resource that is related in any way to them, or a literal or another resource with a label matching a particular text query. In addition, this index also performs subclass inference. Thus, from this stage, in the case of the example queries one

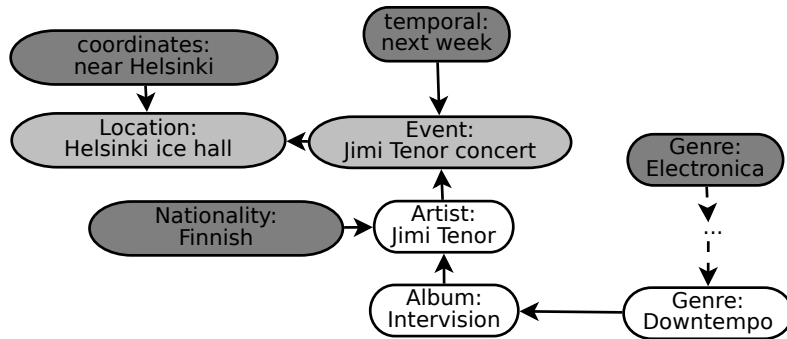


Fig. 2. Mapping to a final search result after keyword and spatiotemporal matching in TravelSampo. Dark grey resources are those returned from matching, while the light grey resources are the final search result.

would already get the artist Jimi Tenor and the album Intervision, in addition to the more direct resource hits.

Then, “mapping” queries are run separately and iteratively for each select query result set. In the example, these would map for example any albums to artists, artists to events and events to locations. After this, the system automatically takes an intersection of the mapped results returned from each select query. A further “filter” query is also run. In TravelSampo, this makes sure that only locations ever make it to the final result set returned.

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

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

6 Contributions

While still a work in progress, the TravelSampo system already demonstrates the potential for a much richer way of searching for points of interest. In developing the system, multiple issues were identified.

Firstly, locations may be of interest not only through their immediate properties, but through quite long chains of associations. Secondly, it is hard to isolate points of interest from other general locations.

In processing actual databases for use in the TravelSampo system, the lack of machine-processable content keywords in most currently available datasets was

identified to be a major problem. In the TravelSampo system, this was addressed by integrating state of the art information extraction tools into the system.

In order to enhance precision and recall in searching the heterogeneous datasets, key class ontology level reference resources in the TravelSampo system such as point of interest types were mapped to each other by hand. However, another requisite part of an integration architecture such as TravelSampo is still the support for iterative, automatic mapping of the instances and keyword concepts in the different datasets pouring in, sometimes dynamically each day. An equally important feature is the ability of human editors to correct these mappings.

Finally, the TravelSampo system and the datasets loaded into it highlight the complexity of queries needed to cater to complex needs, while demonstrating that answering such queries efficiently even on massive data sources is still quite possible.

Acknowledgements This research is part of the Semantic Ubiquitous Services Project (SUBI) 2009–2011, funded by the Finnish Funding Agency for Technology and Innovation (Tekes) and a consortium of 18 companies and public organizations, pre-eminently Turku – European Capital of Culture 2011. Some work has been funded also by the Finnish Cultural Foundation.

References

1. Auer, S., Lehmann, J., Hellmann, S.: LinkedGeoData: Adding a spatial dimension to the web of data. In: The Semantic Web-ISWC 2009. pp. 731–746. Springer (2009), <http://www.springerlink.com/index/j63221026432x374.pdf>
2. Becker, C., Bizer, C.: DBpedia Mobile : A Location-Enabled Linked Data Browser. In: Proceedings of the 1st Workshop about Linked Data on the Web (LDOW2008). Beijing (2008)
3. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: 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), <http://www.sciencedirect.com/science/article/B758F-4WS9BS0-1/2/83cd58f9b584b76ccaa85cda59cca3a2>
4. Choi, N., Song, I.Y., Han, H.: A survey on ontology mapping. SIGMOD Rec. 35, 34–41 (September 2006), <http://doi.acm.org/10.1145/1168092.1168097>
5. Hyvönen, E.: Developing and using a national cross-domain semantic web infrastructure. In: Sheu, P., Yu, H., Ramamoorthy, C.V., Joshi, A.K., Zadeh, L.A. (eds.) Semantic Computing. IEEE Wiley - IEEE Press (May 2010)
6. Hyvönen, E., Mäkelä, E., Kauppinen, T., Alm, O., Kurki, J., Ruotsalo, T., Seppälä, K., Takala, J., Puputti, K., Kuittinen, H., Viljanen, K., Tuominen, J., Palonen, T., Frosterus, M., Sinkkilä, R., Paakkarinen, P., Laitio, J., Nyberg, K.: CultureSampo – Finnish culture on the semantic web 2.0. Thematic perspectives for the end-user. In: Proceedings, Museums and the Web 2009, Indianapolis, USA (April 15–8 2009)
7. Hyvönen, E., Tuominen, J., Kauppinen, T., Väätäinen, J.: Representing and utilizing changing historical places as an ontology time series. In: Ashish, N., Sheth, A. (eds.) Geospatial Semantics and Semantic Web: Foundations, Algorithms, and Applications. Springer-Verlag (2011, forth-coming)

8. Kalfoglou, Y., Schorlemmer, M.: Ontology mapping: the state of the art. *Knowl. Eng. Rev.* 18, 1–31 (January 2003), <http://portal.acm.org/citation.cfm?id=975027.975028>
9. Kurki, J., Hyvönen, E.: Collaborative metadata editor integrated with ontology services and faceted portals. In: Workshop on Ontology Repositories and Editors for the Semantic Web (ORES 2010), the Extended Semantic Web Conference ESWC 2010, Heraklion, Greece. CEUR Workshop Proceedings (June 2010)
10. Medelyan, O.: Human-competitive automatic topic indexing. Ph.D. thesis, University of Waikato, Department of Computer Science (2009)
11. Ruotsalo, T., Mäkelä, E., Kauppinen, T., Hyvönen, E., Haav, K., Rantala, V., Frosterus, M., Dokoohaki, N., Matskin, M.: 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 (2009)
12. Suominen, O., Hyvönen, E.: Expressing and Aggregating Rich Event Descriptions. In: Proceedings of the 6th Workshop on Scripting and Development on the Semantic Web (2010)
13. Wilson, M., Russell, A., Smith, D.A., Owens, A., Schraefel, M.: mSpace mobile: A mobile application for the semantic web. In: Proceedings of the ISWC 2005 End User Semantic Web Interaction Workshop (2005), <http://eprints.ecs.soton.ac.uk/11101>