Distributed Semantic Content Creation and Publication for Cultural Heritage Legacy Systems

Kim Viljanen, Jouni Tuominen, Teppo Känsälä, and Eero Hyvönen

Semantic Computing Research Group
Helsinki University of Technology, Department of Media Technology
and University of Helsinki, Department of Computer Science
P.O. Box 5500, 02015 TKK, Finland
http://www.seco.tkk.fi/ email: first.last@tkk.fi

***Abstract*—Cultural heritage is by nature strongly interlinked, e.g. thematically and historically, but at the same time distributed in heterogeneous collections of different memory organizations at different locations. In order to provide the end-users with aggregated homogeneous views to distributed heterogeneous contents, semantic portals have been created successfully based on metadata and shared (or aligned) ontologies. This paper discusses two problems encountered in such a distributed semantic content creation environment. First, during the content creation work, how could a publisher start using shared ontologies in legacy cataloguing and annotation systems that do not support ontologies. Second, during content publication, how could a publisher re-use the aggregated content in its own legacy publication system, e.g., on the ordinary web pages of a museum or in a collection browser. As a solution, we present the ONKI Ontology Server for adding shared ontological annotation functionalities to legacy cataloguing systems in a practical, cost-efficient and lightweight way. For distributed publishing of the aggregated semantic portal services, we introduce the lightweight mash-up web widget components called "floatlets". A major idea behind both the ONKI functionalities and floatlets is that they can be easily integrated with legacy systems on the user interface level, in the same spirit as e.g. Google Maps.**

I. Introduction

In traditional web publishing, content creators publish web pages and link them together independently from each other. Content management systems (CMS) and portals are used to aggregate related material within one site, and to provide local search and linking services. Search engines are used to provide content aggregation services on the global cross-site level. Linking web pages between sites is usually done manually.

A major goal behind the cultural semantic portals MuseumFinland [1] and CultureSampo [2] is to create a national distributed semantic publishing channel of cultural content on the Semantic Web[1] [3]. The general idea of such a channel is depicted in figure 1 of, where the *content providers* on the left produce metadata about collection items, documents, and other resources of interest along their organizational interests as before for their own purposes ("primary applications" in the figure). Selected content is then harvested into a global knowledge base (center of the figure) to be reused in "secondary applications", such as MuseumFinland or CultureSampo for end-users on the right side of the figure. However, the idea is that the content publishers as well as other external organizations could also by themselves reuse the semantic content cost-effectively, by using the services of the portals in their own web portals. The figure depicts an enhanced portal "Portal 2" in which the content of the primary application is enriched by, e.g., semantic recommendation links to content pages in a semantic portal or by its search services. The *end-users* on the right are provided with global semantic search and browsing services to the semantic portals based on contents aggregated from different heterogeneous collections, being created and maintained at different memory organizations located at different places. At the same time, the enriched services of the primary applications can be used.

This paper addresses two major practical problems in this kind of distributed content creation, aggregation, and publishing model on the Semantic Web. Firstly, when harvesting distributed contents from the memory organizations the metadata cannot be made interoperable with that of other organizations unless it is annotated by using shared metadata schemas and ontologies (or by aligning them). The metadata in cultural collections such as museum databases are often syntactically heterogeneous, contain typos, and are semantically ambiguous based on different vocabularies [4]. This results in lots of tedious syntactic correction, semantic disambiguation, and ontology mapping work when making the contents semantically interoperable. Secondly, when using the semantic services of the aggregated knowledge base and portals in other primary applications, some kind of easy to integrate web service API is needed.

In both cases, our solution approach is to use lightweight mash-up widgets that can be used in legacy and other applications, such as museum cataloging systems or museum CMSs cost-efficiently on the HTML-level as ready-to-use functionalities, with minimal changes in existing systems. The idea is related to Web 2.0 systems such as Google Maps[2] and AdSense[3] but applied to ontology services and semantic service publication.

[1] http://www.w3.org/2001/sw/

[2] http://maps.google.com

[3] http://www.google.com/adsense/

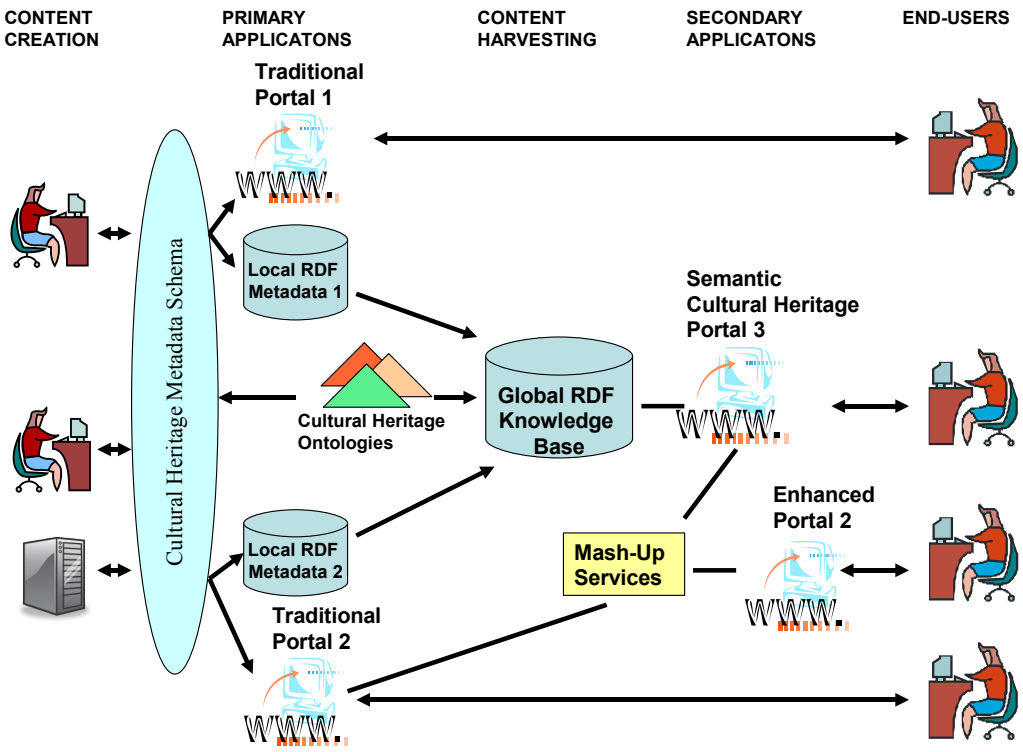

Fig. 1. An overview of a distributed content creation process on the semantic web.

In the following, we discuss the problem of adding ontological annotation functionalities to legacy cataloguing systems and suggest a mash-up approach based on the ONKI Ontology Server [5]. We then discuss the inference and delivery of semantic links as lightweight mash-up web widget components called "floatlets" [6], which can be easily added to each collection owner's legacy website, or to other applications.

II. Distributed Semantic Content Creation

The traditional annotation process for e.g. collection items in a museum cataloging system consists of identifying the correct terms which describe the current item to be annotated, and adding these terms to the item's metadata record. The properties used for content descriptions are specified in the metadata schema. The metadata schema can be based on some metadata standard, e.g. Dublin Core Metadata Element Set[4], which consists of 15 core which contains properties such as Title, Creator and Subject. The use of a shared metadata schema enables only syntactic interoperability and is not enough for our purposes. In addition, the meanings of the elements values, such "bank" in the dc:Subject field, have to be disambiguated leading to semantic interoperability on the semantic web [4]. Instead of using ambiguous human readable terms the metadata should be described using unambiguous ontological concepts, i.e. URIs identifying the concepts.

Typically the functionality to support the annotation process (e.g. concept searching and browsing) is implemented individually in every cataloging system. We propose a mash-up approach to integrate these functionalities into the system cost-efficiently. This idea has been implemented in the ONKI Ontology Server [5]. Because the required functionality is quite general, it can be implemented once in the centralized ontology service and be re-used by different organizations through the web. To maximize the benefits of the integration we provide the functionalities as a ready-to-use user interface component, web widget.

The general idea of the proposed mash-up approach is to provide the developer with a widget that can utilize ONKI services with minimal changes in the legacy system. In the case of an HTML-based legacy system, just a few lines of JavaScript code need to be added on the HTML page. In the case of other user interface technologies, the Web Service[5] interface can be used. The widget solves the problem of fetching correct concept URIs into the application or a database; the actual usage of the acquired semantically correct data is in the responsibility of the target application. This kind of a simple way for getting concept URIs is crucial e.g. in various content creation systems for the semantic web, such as [4], [3].

[4]http://dublincore.org/documents/dces/

[5]http://www.w3.org/TR/ws-arch/

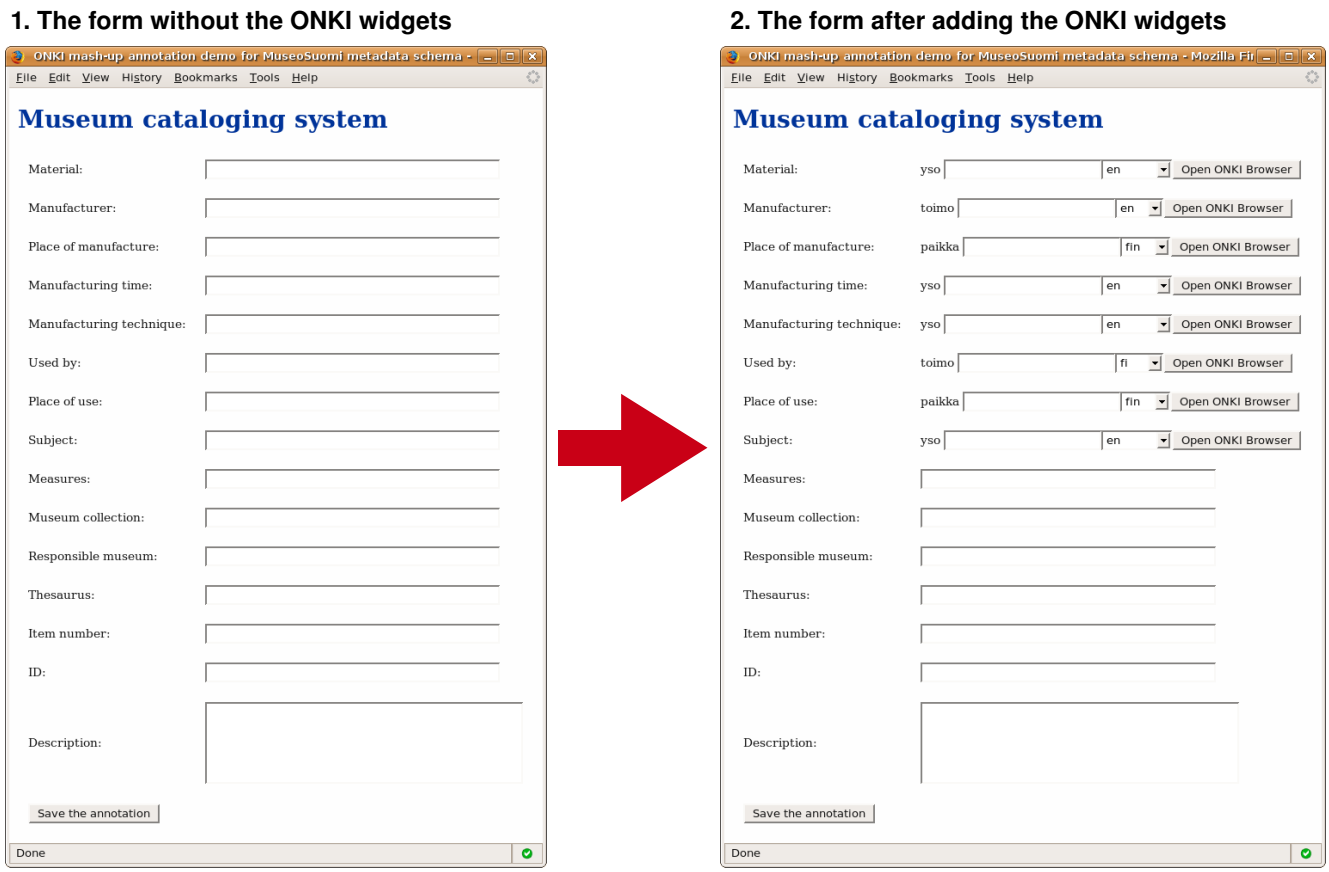

Fig. 2. A museum cataloguing system before and after integrating the ONKI widgets.

The web widget can be integrated into a legacy system by adding the following lines of HTML/JavaScript code into the HTML page.

```
1) <script type="text/javascript"
  src="http://www.yso.fi/onki.js"></script>

2) <input id="dc:subject"
   onkeyup="onki['yso'].search()"/>
```

The code line 1) is used to load the needed ONKI library files and is typically added to the HEAD section of the HTML page. The code line 2) is added to the BODY section of the HTML page to the locations where the ONKI widget component is desired. The string "yso" in the code line 2) refers to the ontology server instance used in the search.

For demonstration purposes, we created a simple web form (see part 1 of figure 2) presenting the MuseumFinland [1] metadata fields. The web form represents a legacy museum cataloguing system. The user interface of the application consists of simple HTML text input fields which are used to describe various properties of the item to be annotated. The fields include e.g. the material of the item, the manufacturer, the place of manufacture and the subject. The museum worker using the system fills the fields with terms from some controlled vocabulary or with free text.

When the museum using the presented system decides to start producing semantic content to achieve the interoperability between the various museum collections, the need to use concepts (i.e. URIs) instead of terms emerges. To support this, the ONKI Concept Search Widget can be utilized. In part 2 of figure 2 the widget is integrated into the application. Every field that is intended to be filled with URIs, is hooked to the ONKI server.

The input fields can be hooked to different ontologies based on the characteristics of the field. For example the ONKI search component in the subject field is configured to search concepts in the Finnish General Upper Ontology YSO[6], the manufacturer field is used for fetching concepts from the Actor Ontology TOIMO[7] and the place of manufacture from the Finnish Geo-ontology SUO[8]. The material field is configured to search concepts in the material sub-branch (the concepts which are subclasses of the concept "material") of YSO.

The ONKI web widget is illustrated in figure 3. When typing a search string to the search field of the mash-up component, the system dynamically performs a query after each input character to the ONKI server, which returns the

[6]http://www.seco.tkk.fi/ontologies/yso/
[7]http://www.seco.tkk.fi/ontologies/toimo/
[8]http://www.seco.tkk.fi/ontologies/suo/

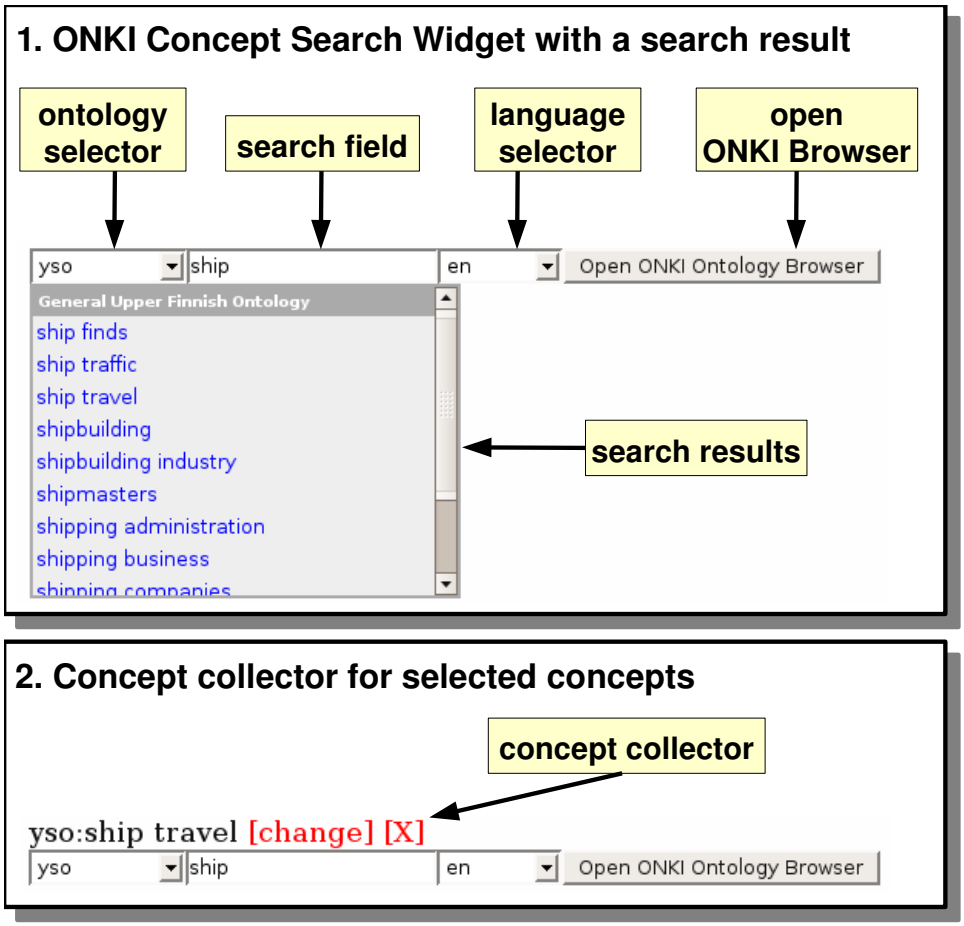

Fig. 3. ONKI Concept Search Widget.

concepts whose labels match the string, given the desired language selection. The ontology to be queried can be changed by using the ontology selector. The results of the query are shown in the web widget's result list below the input field. The desired concepts can be selected from the results. When selected, the concept's URI and label are fetched into the target application. In case of a legacy application, which is not capable of handling URIs, only the labels of concepts can be fetched.

Concepts can also be searched, browsed and fetched with the full screen ontology browser which has been implemented as ONKI-SKOS Browser [5] for lightweight, thesauri-like ontologies and ONKI-Geo Browser [7] for geographical ontologies.

When the desired concepts are fetched, they are stored in a concept collector. The widget provides a default concept collector, but it can also be implemented in the target application. The default concept collector shows the fetched concepts in the widget's user interface, and also stores them in hidden input fields. When the form is submitted, the hidden input fields can be processed by the target application. This way the URIs of the annotation concepts can be transferred, e.g., into the database of the application.

III. Distributed Publishing

Semantic metadata repositories aggregated from distributed content collections can be used for creating services such as semantic link portals [1], [8]. For the end-users, such portals give a global view [9] on the available content in the domain of the portal, not restricted by the (local) view of any single content producer. For the original content producers the main benefit of making their content available in global metadata repositories is to promote their content to wider audiences. However, the content producers typically can not enhance their *own* services, e.g. a museum web site or a collection browser, even though the content of the global repository creates new unique possibilities for e.g. proportioning their local content to the content available elsewhere.

To create more incentives and benefits for the original content producers to publish their content in centralized metadata repositories, we introduce the idea of *distributed publishing* where the global repository provides services that can be used for extending the original content producers' applications and websites based with new global view functionalities. Current solutions for distributed publishing on the web can be broadly divided to 1) automatical statistical content based linking and

2) link feeds published e.g. using RSS[9] or OAI-MHP[10].

The analysis of the textual and link content of the target site can be used for linking the current web page automatically to related ads, as demonstrated by the Google AdSense advertisements, based on the Google technology [10]. These ads can be easily added to any website for getting advertisement revenues using a lightweight mash-up approach that requires no programming skills from the website administrator. Automatic analysis in AdSense has the drawback, that the links may not be semantically related to the content due to misinterpretations of the content. For example, a page containing information about river banks might be linked falsely to financial services (i.e. banks). To provide high quality linking, the link requester should be able to define exactly the subject of the item to get semantically relevant links to other websites. Another problem of textual analysis in cultural content linking is that collection items, such as paintings or photos, may not have much textual metadata describing the content, which makes it more difficult to link related content with each other.

Content item metadata feeds presented e.g. as RSS feeds or using the OAI-MHP protocol are used for publishing a list of all items in a certain collection, e.g., daily news headings or the books in a certain library. By frequently reloading these feeds into an external system, this system can automatically maintain an updated view of the other system's content and e.g. publish the latest information as links on its own website. A problems with metadata feeds is that they require adding new software to the website before links can be shown to the visitors. Also the content presented in the feeds can typically not be configured semantically, which means that the feeds contain usually too general or numerous items to be added to a specific web page presenting e.g. one item in a museum collection.

As the implementation for distributed publishing of semantic web content, we propose the idea of semantic mash-up components, *floatlets*. A floatlet is used for injecting semantic portal functionalities, such as searching and browsing, to existing web pages by publishing the semantic portal functionality as a ready to use mash-up component in the same fashion as e.g. Google Maps and Google AdSense can be connected to any web page. With a floatlet, the host page can be made connected to semantically relevant other content, based on the global metadata repository with knowledge about available content and it's semantical relations.

Search floatlet. The search functionalities typically available in semantic portals may be published as floatlets. This means that the website administrator who adds a floatlet to their site can configure the floatlet to show the results of a semantic search, e.g. faceted search to the centralized metadata repository. For example, when searching for generic ontological concepts such as "chair" or "painting", the floatlet shows the related individuals (instances), such as the Napoleon's coronation chair or the Mona Lisa painting, originating (potentially)

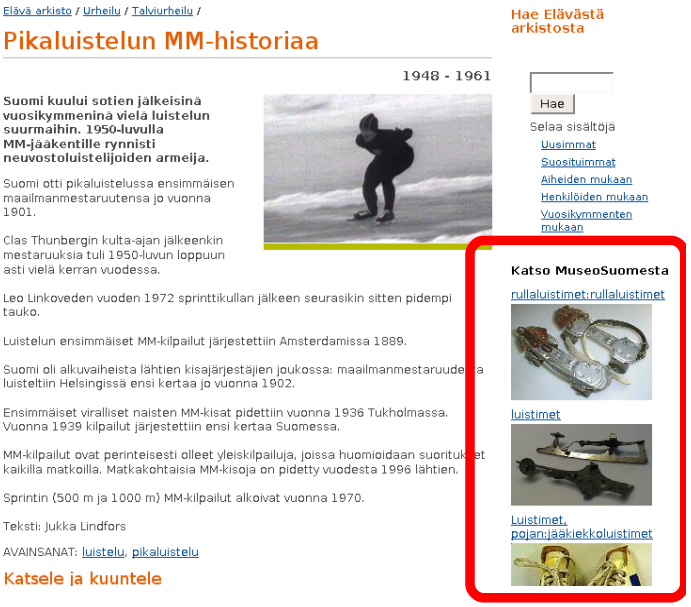

Fig. 4. A floatlet (encircled) displays links to MuseumFinland.

from distinct museums and museum portals. Depending on the capabilities of the underlying search engine, more complex queries such as Boolean queries and graph queries may be supported.

Recommender floatlet. Recommender floatlets provide relevant links to the currently viewed web page based on metadata about the current content which enables cross-portal browsing of semantically related content. For example, when viewing a table manufactured by a certain company, the floatlet may depict a chair manufactured by the same company located elsewhere [11]. The idea of the recommender floatlets is related to Google AdSense that enrich web pages with advertisement links, based on the global Google search indices of web pages. However, in floatlets the returned links are based on explicit ontological metadata of the content, harvested to the global repository, and rule-based reasoning is used for inferring the relevant links [11], [1]. The rule-based approach allows the web developer to specify in detail what information will be linked and why, either manually or based on the metadata of the current web page.

To demonstrate the idea, we have developed a floatlet interface to the MuseumFinland portal[11]. This floatlet can be used by other websites to get relevant links to MuseumFinland. For example, figure 4 shows how a web page of the Finnish Broadcasting Company's video archive[12] has been semantically linked based on metadata with relevant content in MuseumFinland. In the example, the current video is about the history of speed skating, which has also been described in its metadata. Based on this information, the floatlet is able to query for old skates from MuseumFinland. By clicking on

[9]http://www.rssboard.org/rss-specification/
[10]http://www.openarchives.org/OAI/openarchivesprotocol.html
[11]http://www.museosuomi.fi
[12]http://www.yle.fi/elavaarkisto/

the floatlet links, the skates can be examined in more detail in MuseumFinland.

The example floatlet can be added to the webpage by adding the following two lines of JavaScript:

```
<script type='text/javascript'
  src='http://demo.seco.tkk.fi/leijuke/
    semlink_and_dojo_all.js'></script>

<div dojoType='seco:Recommend'
  concepts='http://www.cs.helsinki.fi/group/
  seco/ns/2004/03/30-masa#skates'>Loading</div>
```

The concept URI (skates) refers to the Cultural Heritage Ontology MAO used in MuseumFinland. Instead of URIs, also keywords may be used as parameters.

IV. Discussion

This paper considered the problem of supporting legacy systems in distributed content creation and publishing for the semantic web. As a practical solution we presented the ONKI ontology server with a *concept search widget* that can be easily added to cataloguing and metadata systems. We also presented the *floatlets* for publishing semantic portal functionalities in other web sites, especially to be used by the original content producers of a metadata repository. We argue, that the interest of the original content producers to participate in the Semantic Web can be enhanced by 1) making the creation of high quality semantic metadata in content creators' legacy systems as easy as possible and 2) by making it possible to benefit from the harvested metadata in content creators' *own* applications and web sites. The possibility to enhance the content producers' own applications and web pages with floatlets (based on the global semantic portal) might be an important incentive to participate in the distributed publishing network, especially if additional work is required to create interoperable metadata.

Our technical approach in both the ONKI concept search widget and the floatlets was to create a centralized semantic web server which provides in addition to ordinary browser based user interfaces (the ONKI concept browser [5], respectively semantic portal such as MuseumFinland [1]) also lightweight mash-up widgets that can be easily added to existing legacy applications. Based on our experiences, the approach seems valid and effective.

Future work includes evaluating the proposed content creation and publishing methods in real production environments to gain more knowledge about the applicability and potential problems of the solutions. As a result of using the ONKI concept search widget, legacy systems contain semantically annotated content but lack the semantic search and other functionalities for using the content. We envision the possibility to provide also such functionalities as mash-up services for the legacy systems.

Acknowledgments

We thank Ville Komulainen for his work on the original ONKI server. This work is a part of the National Semantic Web Ontology project in Finland[13] (FinnONTO), funded mainly by the National Technology and Innovation Agency Tekes and 36 private, public and non-governmental organizations.

References

[1] E. Hyvönen, E. Mäkelä, M. Salminen, A. Valo, K. Viljanen, S. Saarela, M. Junnila, and S. Kettula, "Museumfinland – finnish museums on the semantic web," *Journal of Web Semantics*, vol. 3, no. 2, p. 25, 2005.

[2] E. Hyvönen, T. Ruotsalo, T. Häggström, M. Salminen, M. Junnila, M. Virkkilä, M. Haaramo, T. Kauppinen, E. Mäkelä, and K. Viljanen, "CultureSampo—Finnish culture on the semantic web. The vision and first results," in *Information Technology for the Virtual Museum*, K. Robering, Ed. LIT Verlag, Berlin, 2007.

[3] E. Hyvönen, K. Viljanen, and O. Suominen, "HealthFinland—Finnish health information on the semantic web," in *Proceedings of the 6th International Semantic Web Conference (ISWC 2007), Busan, Korea*. Springer-Verlag, Nov 2007.

[4] E. Hyvönen, M. Salminen, S. Kettula, and M. Junnila, "A content creation process for the Semantic Web," 2004, proceeding of OntoLex 2004: Ontologies and Lexical Resources in Distributed Environments, Lisbon, Portugal.

[5] K. Viljanen, J. Tuominen, and E. Hyvönen, "ONKI ontology server—extending legacy systems with ontology mash-up services," November 2007, submitted for review. http://www.seco.tkk.fi/publications/.

[6] E. Mäkelä, K. Viljanen, O. Alm, J. Tuominen, O. Valkeapää, T. Kauppinen, J. Kurki, R. Sinkkilä, T. Känsälä, R. Lindroos, O. Suominen, T. Ruotsalo, and E. Hyvönen, "Enabling the semantic web with ready-to-use web widgets," in *Proceedings of the First Industrial Results of Semantic Technologies Workshop, ISWC2007*, 2007.

[7] E. Hyvönen, R. Lindroos, T. Kauppinen, and R. Henriksson, "An ontology service for geographical content," in *Poster Proceedings of the 6th International Semantic Web Conference (ISWC/ASWC 2007), Busan, Korea*, Nov 2007.

[8] M. Hildebrand, J. van Ossenbruggen, and L. Hardman, "/facet: A browser for heterogeneous semantic web repositories," in *Proc.s of 5th International Semantic Web Conference (ISWC 2006)*. Springer-Verlag, 2006, pp. 272–285.

[9] A. Calí, G. D. Giacomo, and M. Lenzerini, "Models for information integration: Turning local-as-view into global-as-view," in *Proc. of Int. Workshop on Foundations of Models for Information Integration (10th Workshop, Foundations of Models and Languages for Data and Objects)*, 2001. [Online]. Available: citeseer.ist.psu.edu/495550.html

[10] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Stanford Digital Library Technologies Project, Tech. Rep., 1998. [Online]. Available: citeseer.ist.psu.edu/page98pagerank.html

[11] K. Viljanen, T. Känsälä, E. Hyvönen, and E. Mäkelä, "Ontodella - a projection and linking service for semantic web applications," in *Proceedings of the 17th International Conference on Database and Expert Systems Applications (DEXA 2006), Krakow, Poland*. IEEE, September 4-8 2006, pp. 370–376.

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