A Semantic Layer for Urban Resilience Content Management

Ilkka Niskanen, Mervi Murtonen
Technical Research Centre of Finland
Oulu/Tampere, Finland
Ilkka.Niskanen@vtt.fi,
Mervi.Murtonen@vtt.fi

Fiona Browne, Peadar Davis
School of Computing and Mathematics
Ulster University
Jordanstown, Northern Ireland, UK
f.browne@ulster.ac.uk
pt.davis@ulster.ac.uk

Francesco Pantisano
Smart System Infrastructures
Finmeccanica Company
Genova, Italy
francesco.pantisano@finmeccanica.co

***Abstract*— Content Management refers to the process of gaining control over the creation and distribution of information and functionality. Although there are several content management systems available they often fail in addressing the context specific needs of end-users. To enable more task specific and personalized support we present a semantic content management solution developed for the domain of urban resilience. The introduced semantic layer is built on top of an existing content management system and by utilizing domain specific annotation and categorization it facilitates the management of heterogeneous and large content repository. In addition, the enhanced semantic intelligence allows better understanding of content items, linkages between unstructured information and tools, and provides more sophisticated answers to users' various needs.**

Keywords- content management; semantic technologies; heterogeneous data repository

I. INTRODUCTION

The field of Content Management (CM) refers to the process of gaining control over the creation and distribution of information and functionality. Concisely, an effective Content Management System (CMS) aims at getting the right information to the right people in the right way. Usually CM is divided into three main phases namely collecting, managing, and publishing of content. The collection phase encompasses the creating or acquiring information from an existing source. This is then aggregated into a CMS by editing it, segmenting it into components, and adding appropriate metadata. The managing phase includes creating a repository that consists of database containing content components and administrative data (data on the system's users, for example). Finally, in the publishing stage the content is made available for the target audience by extracting components out of the repository and releasing the content for use in the most appropriate way. [1]

Currently, there are several commercial and open-source technologies available that are applied to address different content management needs across various industries including healthcare [12], and education [15], for example. However, the standard versions of the existing solutions are not always capable of supporting end-users in their specified context to reach their particular goals in an effective, efficient and satisfactory way [2]. For instance, the included content retrieval mechanisms are often implemented using traditional keyword based search engines that are not adapted to serve any task specific needs [3][4][5].

One of the main issues to be resolved is how to convert existing and new content that can be understood by humans into semantically-enriched content that can be understood by machines [6]. The human-readable and unstructured content is usually difficult to automatically process, relate and categorize, which hinders the ability to extract value from it [7]. Additionally, it results in the restriction of development of more intelligent search mechanisms [6]. To address some of the above described deficiencies, semantic technologies are being increasingly used in CM. In particular, the utilization of domain specific vocabularies and taxonomies in content analysis enables accurate extraction of meaningful information, and supports task-specific browsing and retrieval requirements compared to traditional approaches [6]. Furthermore, semantic technologies facilitate creating machine-readable content metadata descriptions, which allows, for example, software agents to automatically accomplish complex tasks using that data. Moreover, semantically enhanced metadata helps search engines to better understand what they are indexing and providing more accurate results to the users [9].

The HARMONISE platform, developed in the FP7 EU HARMONISE [17] project is a domain specific CMS that provides information and tools for security-driven urban resilience in large-scale infrastructure offering a holistic view to urban resilience. A database contained by the system manages an extensive set of heterogeneous material that comes in different forms including tools, design guidance and specifications. The platform aims at serving as a 'one-stop-shop' for resilience information and guidance and it contains a wealth of information and tools specifically designed to aid built environment professionals. While the platform and the hosted toolkit are aimed to be used by a variety of potential end-users from planners and urban designers to construction teams, building security personnel and service managers, the specialized problem domain and heterogeneous content repository poses significant challenges for users to effectively retrieve information to accomplish their tasks and goals.

In this paper the Semantic Layer for the HARMONISE (SLH) approach is introduced. The SLH is a semantic content management solution developed to address many of the above discussed challenges related to domain specific

 ISBN: 978-1-61208-472-5

content management. It is implemented on top of the HARMONISE platform and it aims at offering more task specific and personalized content management support for end-users. Additionally, by utilizing domain specific annotation and categorization of content the SLH facilitates the management of heterogeneous and large content repository hosted by the HARMONISE platform.

The semantic information modelling allows better understanding of platform content, linkages between unstructured information and tools, and more sophisticated answers to users' various needs. Moreover, the semantic knowledge representations created by the layer help end-users to combine different data fragments and produce new implicit knowledge from existing data sets. Finally, by utilizing Linked Data [18] technologies the SLH fosters interoperability and improves shared understanding of key information elements. The utilization of interconnected and multidisciplinary knowledge bases of the Linked Data cloud also enables applying the solution in other problem areas such as health care or education.

The rest of paper is organized as follows. Section II provides a through description of the HARMONISE platform and its application area. In Section III the architecture and different components of the SLH are described. Section IV provides a Use Case example demonstrating the functionality of the SLH. Finally, Section V concludes the paper.

II. THE HARMONISE CONTENT MANAGEMENT PLATFORM

At present there exist a number of content management systems that enable publishing, managing and organizing electronic documents. For example, Drupal [19] and WordPress [20] are well-known, general-purpose CM solutions providing such basic CM features such as user profile management, database administration, metadata management, and content search and navigation functionalities [2]. These tools provide functionality to create and edit a website's content often with easy-to-use templates for digital media content publishing.

As stated above, the HARMONISE platform is a CMS specifically tailored for the domain of urban resilience. The system provides information and tools for security-driven urban resilience in large-scale infrastructure and contains a variety of interactive elements allowing users to both import and export data to and from the platform and personalize the platform to their own needs. The core functionalities of the HARMONISE platform are implemented using ASP.NET web application framework and it utilizes Microsoft SQL 2012 database to store content items.

An important part of the HARMONISE content management platform is the Thematic Framework [10] that was created to structure information within the platform and to guide end-users through an innovative step-by-step search process. The Thematic Framework is set out in Fig. 1 below.

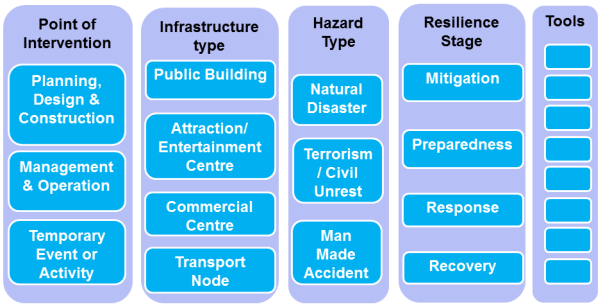

Figure 1. The Thematic Framework (adopted from [10])

By unpacking resilience into a number of key layers the Thematic Framework provides the necessary taxonomy needed for realizing effective domain-specific content annotation and categorizing functionalities, as later discussed. The objective of the domain-specific annotation is to allow users to easily identify and access information and tools within the platform, and to search the platform according to their unique needs or interests.

III. THE SEMANTIC LAYER

As earlier described, the HARMONISE content management platform hosts a large portfolio of urban resilience related content. However, finding relevant information and tools from such a knowledge base with conventional information retrieval methods is usually both tedious and time consuming, and tends to become a challenge as the amount of content increases [6]. Often users have difficulties in grouping together related material or finding the content that best serve their information needs, especially when content is stored in multiple formats [11].

In general, the existing CMSs usually lack consistent and scalable content annotation mechanisms that allow them to deal with the highly heterogeneous domains that information architectures for the modern knowledge society demand [8]. The semantic layer described in this study aims at addressing the above mentioned challenges by integrating semantic data modelling and processing mechanisms to the core HARMONISE platform functionalities. For example, the application of semantic mark-up based tagging of web content enables expressively describing entities found in the content, and relations between them [6]. Moreover, by utilizing the Linked Data Cloud links can be set between different and heterogeneous content elements and therefore connect these elements into a single global data space, which further facilitates interoperability and machine-readable understanding of content [13].

The main features of the SLH are divided to four parts. First, the metadata enrichment part produces information-rich metadata descriptions of the content by enhancing content with relevant semantic metadata. Second, the semantic metadata repository implements the necessary means for storing and accessing the created metadata. The third component of the SLH realizes a semantic search feature. In more detail the search service aims at returning more meaningful search results to the user by utilizing both keyword-based semantic search and "Search by theme"

 ISBN: 978-1-61208-472-5

filtering algorithm that restricts the searchable space by enabling users to select certain categories from the Thematic Framework. The final part, content recommendation, combines information about users' preferences and profile to find a target user neighborhood, and proactively recommends new urban resilience tools/resources that might be of potential interest to him/her. In Fig. 2 the logical architecture of the SLH is represented.

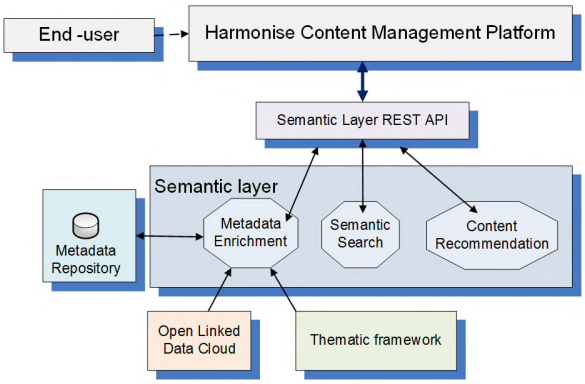

Figure 2. The logical architecture of the SLH

The following sections describe the logical architecture in more detail.

A. Semantic Layer REST API

The Semantic Layer REST API provides the necessary interface for the HARMONISE Platform to interact with the SLH. It enables, for example, to transmit query requests from the platform to the SLH or retrieve content recommendations personalized for a particular user.

B. Metadata Enrichment

The purpose of the Metadata Enrichment service is to produce information-rich metadata descriptions of the content that is uploaded to the HARMONISE platform. Enhancing content with relevant semantic metadata can be very useful for handling large content databases [1]. A key issue in this context is improving the "findability" of content elements (e.g. documents, tools).

The enrichment process is based on tagging. A tag associates semantics to a content item, usually helping the user searching or browsing through content. These tags can be used in order to identify the most important topics, entities, events and other information relevant to that content item. The tagging data is created by analyzing the uploaded content and the metadata manually entered by the user. This information consist e.g. title, keywords, Thematic Framework categories, topics, content types and phrases of natural language text.

In the metadata analysis the following three technologies that provide tagging services are utilized: ONKI [21], DBPedia [22] and OpenCalais [23]. The ONKI and DBPedia knowledge bases provide enrichment of the human defined keywords by utilizing Linked Data reference vocabularies and datasets. The Metadata Enrichment service utilizes the APIs of the above mentioned technologies to search terms that are somehow associated to the entities defined by a user.

The extracted terms fall into three categories: similar, broader and narrower. The similar terms are synonyms to the original entities whereas broader terms can be considered as more general concepts. The narrower terms represent examples of more specific concepts compared to the original entity. Each of the acquired terms contains a Linked Data URI that can be accessed to get more extensive description of that term. By enriching the human defined keywords with additional concepts and Linked Data URIs more comprehensive and machine-readable information about uploaded content items can be generated.

The uploaded content items are also examined using the OpenCalais text analyzer tool. Using such mechanisms as natural language processing and machine learning the tool allows analyzing different text fragments contained by the uploaded content item. As a result, OpenCalais discovers entities (Company, Person etc.), events or facts that are related to the uploaded content element.

In the final part of the metadata enrichment process the metadata elements created by different tools are merged as a single RDF (Resource Description Framework) metadata description and stored to the metadata database.

C. Semantic Metadata Repository

The database technology used for storing the semantic metadata of content is OpenLink Virtuoso [26]. Virtuoso is a relational database solution that is optimized to store RDF data. It provides good performance and extensive query interfaces [16] and was thus selected as the metadata storage to be used in the SLH.

D. Semantic Search

The Semantic Search service aims at producing relevant search results for the user by effectively utilizing the machine-readable RDF metadata descriptions created by the Metadata Enrichment service. Unlike traditional search engines that return a large set of results that may or may not be relevant to the context of the search, the Semantic Search analyses the results and orders them based on their relevancy. Thus, users are emancipated from performing the time-consuming work of browsing through the retrieved results in order to find the content they are looking for.

The Semantic Search service is implemented as a Java web application composed of three main components (see Fig. 3):

- RESTful Web Service: based on Apache CXF framework, it represents the semantic search service front-end. It receives the search queries from the HARMONISE platform and returns the list of search results provided by the underlying components;
- Semantic Search Service Core (SSS Core): component based on Java/Maven project, customized to manage all the core processes (data indexing, content search, content retrieving, results formatting);
- Semantic Search Engine: component based on Apache Solr [24] enterprise search platform, in charge of the indexing and the search processes. When a new content

 ISBN: 978-1-61208-472-5

is uploaded to the HARMONISE platform it reads from the Virtuoso database the data produced by the semantic content enrichment service in order to create the index to query on. When a user submits a query the semantic search engine queries the index in order to find the documents that best match the user request parameters.

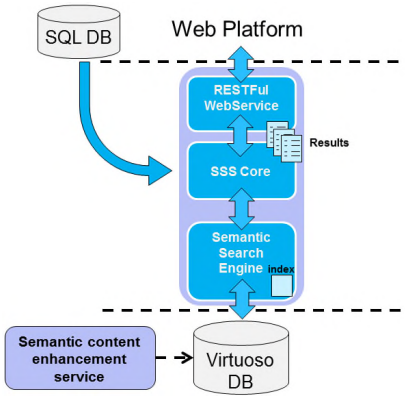

Figure 3. Logical architecture of the semantic search service

The Semantic Search service relies on the Solr search engine [25] in order to search across large amount of content metadata and pull back the most relevant results in the fastest way. Solr is a document storage and retrieval engine, which uses Lucene's inverted index to implement its fast searching capabilities. Unlike a traditional database representation where multiple documents would contain a document ID mapped to some content fields containing all of the words in that document, an inverted index inverts this model and maps each word to all of the documents in which it appears. Solr stores information in its inverted index and queries that index to find matching documents.

In the Semantic Search service, the Solr index is constructed according to the Metadata Repository data structure. In more detail, a sample of the following data fields are encompassed in the index: Id (document identifier on Virtuoso DB); Upload date (date when the document has been uploaded); Topics (list of topics from the Thematic Framework); Permissions (list of user groups allowed to view the document), Description (description of the document) and Tags (list of tags added by the metadata enhancement service).

The search results provided by the Semantic Search service are ranked according to the relevancy scores that measure the similarity between the user query and all of the documents in the index. The results with highest relevancy scores appear first in the search results list.

The scoring model is composed by the following scoring factors:

- Term Frequency: is a measure of how often a particular term appears in a matching document. Given a search query, the greater the term frequency value, the higher the document score.
- Inverse Document Frequency: is a measure of how "rare" a search term is. The rarer a term is across all documents in the index, the higher its contribution to the score.
- Coordination Factor: It is the frequency of the occurrence of query terms that match a document; the greater the occurrence, the higher is the score.
- Field length: the shorter the matching field, the greater the document score. This factor penalizes documents with longer field values.
- Boosting: is the mechanism that allows to assign different weights to those fields that are considered more (or less) important than others.

E. Content Recommendation

Similar to the Semantic Search, the Content Recommendation Service (CRS) is based on semantic modelling of content resources. The aim of the content recommendation service is to improve user experience in terms of the search functionality and the filtering of relevant information through the utilization of collaborative filtering.

The CRS utilizes user profiles which are created and maintained by the HARMONISE platform. The CRS combines information about users' preferences and profile to find a target user neighborhood, and recommend new urban resilience tools/resources that might be of potential interest to him/her. Ordered weighted average and uniform aggregation operators are applied to fuse user information and obtain global degrees of similarity between them. The user profiles contain information about user's preferences and favorite content, for example. It also includes the content item IDs that have been already recommended for that particular user. This information is then utilized when content recommendations are created for different users. An overview of the CRS algorithm is provided in Fig. 4. Fig 4 illustrates how user preference and user profile similarity are fused together along with a weighted sum to provide a ranked list of recommendation tailored to the user.

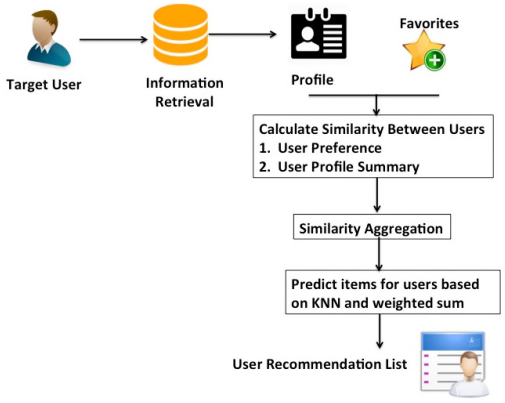

Figure 4. Overview of the CRS Algorithm

The CRS is triggered by the HARMONISE platform through the 'get recommendation' method provided by the Semantic Layer REST API. The ID of the user is transmitted as a method parameter. Once the recommendation service receives the ID, it retrieves the user profile of the user from the database and analyses the information it contains. It

 ISBN: 978-1-61208-472-5

extracts, for example, the topics and research areas the user is interested in. Additionally, the profession, areas of expertise and relevant user groups are retrieved from the user profile. The algorithm then identifies similar users based upon the user profile using the Jaccard [14] index. Similarity between users is also measured by taking into consideration their profile similarity using the ordered weighted sum. Both these measures are fused using a similarity aggregation approach.

The actual recommendation generation process is carried out by comparing the user profile data with the semantic content metadata descriptions. Similarly as in the search algorithm described in the previous section, the content items whose metadata is associated with e.g. terms, topics or research areas as contained by the user profile are included to the initial recommendation results. Of course, the content items that have already been recommended for the user are excluded from the results list. Subsequently, the recommendation results are analyzed using the ranking model introduced by the Semantic Search. Using K-nearest neighbors, the content items that gets the highest score is returned to the platform as the most highly recommended content item.

IV. USE CASE EXAMPLE

The functionality of the SLH is demonstrated with a Use Case example in which a user uploads a document into the HARMONISE content management platform and tries to retrieve it with the search functionality. Additionally, the recommendation service is verified by creating a user profile that is interested in topics relevant to the uploaded content. The content item used in the Use Case example is an electronic manual that presents tools to help assess the performance of buildings and infrastructure against terrorist threats and to rank recommended protective measures. This kind of guidance document is a typical representative of a content item managed by the HARMONISE platform.

Once the user has provided necessary input in the upload form the content description is transmitted to the Metadata Enrichment component that processes the collected data and forms an RDF metadata description of the content. It was noted that the returned semantic content metadata contained five keywords that are enriched with 81 broader or narrower and 26 similar terms. Moreover, the content is annotated with several categories defined by the Thematic Framework.

Once the enriched metadata is stored to the Semantic Metadata Repository, and indexed by the Semantic Search service, it can be tried to be retrieved with the search functionality. The content retrieval is tested with the 'Resilience Search Wizard' feature provided by the SLH. The wizard allows to define keywords and to select those categories from the Thematic Framework that are considered as relevant to the uploaded content. The utilized search parameters are shown in the search wizard screenshot illustrated in Fig. 5.

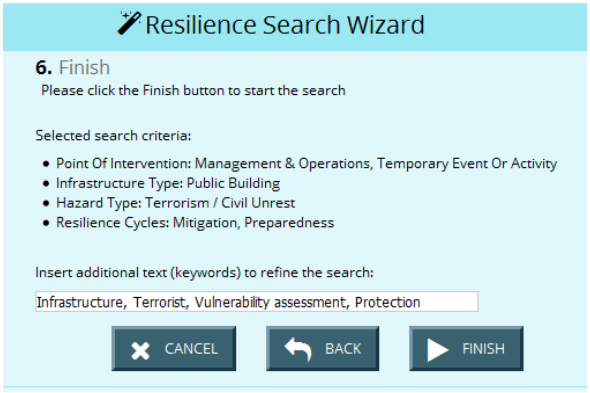

Figure 5. Search parameter definition

As earlier explained, the search functionality is able to sort the results based on their relevancy. Fig. 6 represents the most highly ranked search results returned by the search service. As can be seen, the applied ranking algorithm identified the uploaded electronic manual document as the second relevant search result for the given search query. In total, the search functionality found 24 results with the defined search parameters.

Figure 6. The ranking of search results

In the final phase of the use case example, the Content Recommendation service is tested by creating a user profile and obtaining personalized recommendations. The user profile was created with 6 topics of interests of interest from a total of 13 topics namely: Point of Intervention, Management and Operation, Infrastructure type, Commercial Center, Hazard Type and Man Made Hazard. The user then marked 10 items of favorite content from a total of 156 items in the database. These included content such as "Tools of Regional Governance" and "Flood management in Linares Town". For the first step in the recommendation algorithm, Jaccard index is utilized to compute the degree of similarity between the favourite content and profile information of the user entered and all the users of the HARMONISE system. In the second step, a KNN algorithm is applied to identify the 5 most similar neighbors. Based on neighbor users, we compute for each item not marked as a favorite by the user, a predicted rating. This is used to construct an ordered recommendation list to the target user, which in this case study was a list of 5 recommendations including documents based on "Key issues of Urban Resilience", "Building urban

 ISBN: 978-1-61208-472-5

resilience Details" and "Resilience: how to build resilience in your people and your organization".

V. CONCLUSION

In this work, we introduce a developed semantic content management system for the domain of urban resilience. This system utilizes semantic technologies to manage an extensive set of heterogeneous material that comes in different forms including tools, design guidance documentation and specifications. Moreover, the developed approach enables the creation of machine-understandable and machine-processable descriptions of content items. This has resulted in an improved shared understanding of information elements and interoperability.

The described approach was implemented on top of an existing content management system. With the effective utilization of Linked Data based analysis tools and domain specific content annotation mechanisms it offers task specific and personalized content management support for end-users. The enhanced intelligence has provided better understanding of urban resilience content, linkages between unstructured information and tools, and more sophisticated answers to users' various needs.

With minimal adjustments the introduced semantic layer could be utilized also in other problem domains. For a new CMS to integrate with the semantic layer requires only creating a well described domain specific taxonomy and implementing the technical means for communicating with the provided REST API.

Up to this point, the HARMONISE platform and the SLH have been tested by HARMONISE project partners and other invited domain specialists who have evaluated the system in terms of usability, perceived usefulness and the relevancy of received search and recommendation results. Next, the evaluation process will encompass final tests where the approach will be used in problem area specific case studies with various end user groups.

The future work also includes further refining the HARMONISE platform and the SLH on the basis of the feedback received from the case studies. Additionally, the graphical appearance of the platform's user interface as well as the usability of individual components will be improved.

REFERENCES

[1] B. Boiko, Content management bible, John Wiley & Sons, 2005.

[2] N. Mehta, Choosing an Open Source CMS: Beginner's Guide. Packt Publishing Ltd, 2009.

[3] D. Dicheva and D. Christo, "Leveraging Domain Specificity to Improve Findability in OER Repositories." Research and Advanced Technology for Digital Libraries. Springer Berlin Heidelberg, pp. 466-469, 2013.

[4] S. K. Patel, V. R. Rathod, and S. Parikh, "Joomla, Drupal and WordPress-a statistical comparison of open source CMS." Trendz in Information Sciences and Computing (TISC), 3rd International Conference on. IEEE, 2011.

[5] C. Dorai and S. Venkatesh, "Bridging the semantic gap in content management systems." Media Computing. Springer US, pp. 1-9, 2002.

[6] J. L. Navarro-Galindo and J. Samos, "The FLERSA tool: adding semantics to a web content management system." International Journal of Web Information Systems 8.1: pp. 73-126, 2012.

[7] A. Kohn, F. Bry, and A. Manta, "Semantic search on unstructured data: explicit knowledge through data recycling." Semantic-Enabled Advancements on the Web: Applications Across Industries: Applications Across Industries, 194, 2012.

[8] R. García, J. M. Gimeno, F. Perdrix, R. Gil, and M. Oliva, "The rhizomer semantic content management system", In Emerging Technologies and Information Systems for the Knowledge Society pp. 385-394, Springer Berlin Heidelberg, 2008.

[9] D. R. Karger and D. Ouan, "What would it mean to blog on the semantic web?" The Semantic Web–ISWC 2004. Springer Berlin Heidelberg, pp. 214-228, 2004.

[10] S. Purcell, W. Hynes, J. Coaffee, M. Murtonen, D. Davis, and F. Fiedrich, "The drive for holistic urban resilience" 9th Future Security, Security Research Conference, Berlin Sep. 16- 18, 2014.

[11] A Vailaya, M. A. Figueiredo, A. K. Jain, and H. J. Zhang, "Image classification for content-based indexing". Image Processing, IEEE Transactions on, 10(1), pp. 117-130, 2001.

[12] S. Das, L. Girard, T. Green, L. Weitzman, A. Lewis-Bowen, and T. Clark, "Building biomedical web communities using a semantically aware content management system". Briefings in bioinformatics, 10(2), pp. 129-138, 2009.

[13] M. Hausenblas, "Exploiting linked data to build web applications." IEEE Internet Computing 4, pp. 68-73, 2009.

[14] R. Real and J. M. Vargas, "The probabilistic basis of Jaccard's index of similarity." Systematic biology, pp. 380-385, 1996.

[15] N. W. Y Shao, S. J. H Yang, and A. Sue, "A content management system for adaptive learning environment." Multimedia Software Engineering, Proceedings. Fifth International Symposium on, IEEE, 2003.

[16] O. Erling and I. Mikhailov. "RDF Support in the Virtuoso DBMS", Networked Knowledge-Networked Media. Springer Berlin Heidelberg, pp. 7-24, 2009.

[17] The HARMONISE project (Available online at: http://harmonise.eu/) [accessed: 13.4.2016]

[18] Linked Data (Available online at: http://linkeddata.org/) [accessed: 13.4.2016]

[19] Drupal (Available online at: https://www.drupal.org/) [accessed: 13.4.2016]

[20] WordPress (Available online at: https://wordpress.org/) [accessed: 13.4.2016]

[21] ONKI - Finnish Ontology Library Service (Available online at: http://onki.fi/) [accessed: 13.4.2016]

[22] DBpedia (Available online at: http://dbpedia.org/) [accessed: 13.4.2016]

[23] OpenCalais (Available online at: http://www.opencalais.com/) [accessed: 13.4.2016]

[24] Solr (Available online at: http://lucene.apache.org/solr/) [accessed: 13.4.2016]

[25] Apache Lucene Core (Available online at: https://lucene.apache.org/core/) [accessed: 13.4.2016]

[26] Virtuoso Universal Server (Available online at: http://semanticweb.org/wiki/Virtuoso) [accessed: 13.4.2016]

 ISBN: 978-1-61208-472-5