Semantic web developments in Hungarian public collections in an international perspective

Márton Németh, mnemeth@monguz.hu
Monguz Ltd, Hungary

Libellarium, IX, 2 (2016): 169 – 186.
UDK: 02:004.738.52(439)=111
DOI: http://dx.doi.org/10.15291/libellarium.v9i2.262
Professional paper

Abstract

This article first describes the international context of Hungarian semantic web projects. It then introduces two semantic web-based implementation projects implemented at the Hungarian National Library and the Petőfi Literary Museum. It also describes the results of ALIADA projects helping public collections to appear on the semantic web. Some future development plans and the basic concept of the Hungarian National Namespace are presented next. Finally, it introduces the semantic functions of the HTML5 document markup standard that can offer useful solutions for public collections.

KEYWORDS: Semantic Web, National Széchényi Library, Petőfi Literary Museum, ALIADA project, HTML5 microformats

Introduction

Related to a PhD project at the University of Debrecen, Faculty of Informatics, two semantic web-based implementation projects are introduced in this article. The first project is implemented at the Hungarian National Library (as the pioneering project in this field in Hungary) and the second one at the Petőfi Literary Museum that currently builds up its own database from authority names in order to publish their data on the semantic web. Some currently developed tools are described through the current results of the ALIADA project that helps public collections to appear on the semantic web. Some solutions are also introduced that can help public collections to publish their contents on the semantic web by the help of semantic functions of the HTML 5 document markup standard.

More and more libraries, museums and archives want to publish their standardized datasets on the semantic web. In order to reach this goal, they have to build up semantic ontologies. A semantic ontology is an explicit specification of a conceptualization. RDF/OWL language is appearing as a representation ways of ontologies. Metadata description and cataloguing is appearing in RDF/XML language. Other types of standard data inputs (like Dublin Core, LIDO) must be converted to that format. Namespaces can identify the different kind of data inputs that are appearing in RDF/XML environment. Thesauri and authority data can also appear in semantic web environments in this way (VIAF, FOAF, SKOS). Data must be published as linked open data in order to build-up standard connections with other standard RDF/XML based datasets.

In this paper, after describing the international context of the Hungarian semantic web projects, a semantic web project of the Hungarian National Széchényi Library (NSZL) is introduced. NSZL built up an ontology and published their catalogue and authority data as linked open data on the semantic web (with the help of an own name space, SKOS and VIAF). They also generated a SPARQL-endpoint in order to put their semantic datasets on the cloud.

A recent project of the Petőfi Literary Museum focuses on name authority record. They are converting around 620,000 authority record into RDF-XML format. They would like to publish their data in VIAF (worldwide semantic virtual authority file database) and make them accessible on the OCLC WorldCat semantic environment. This database will be the second database followed by the Hungarian National Library that will appear in VIAF. They also plan to build up an own triple-store in collaboration with other Hungarian libraries and museums in order to enrich their own semantic datasets with external semantic links. In this context some tools need to be introduced which are developed in the framework of the ALIADA project with an international collaboration in order to offer a complete environment to public collection to help them appear on the semantic web with their own triple-store and datasets. Some major projects are described in the semantic field in Hungary that can be implemented possibly in the future. The concept of the Hungarian National Namespace that can set all the institutional efforts from public collections to a common framework is highlighted. Last but not least, the use of microformats (based on HTML 5 standard) is described to put semantic markup data elements into full-text online content resources. That can be also really relevant for public collections (especially for libraries) in order to provide new semantic datasets based on their existing full-text online resources.

International context of the Hungarian semantic web projects

In this section of the paper some international service models are being described that had a great effect to the implementation of Hungarian semantic web projects.

The LIBRIS Project in the Swedish National Library

The Swedish National Library was among the first that released its whole catalogue called LIBRIS as linked data including authority data, describing persons, organizations and subject headings. Links were added to external resources such as those described by Library of Congress Subject Headings (LCSH), Wikipedia and the Virtual International Authority File (VIAF). An important development step is that LIBRIS has become a part of quickly expanding graph of metadata generated by a number of entities, mostly outside the GLAM (Galleries, Archives, Museums, and Libraries) sector. This move inspired a lot of interest especially from other government entities and other organizations that wanted to either link to or download parts of the authority data. It has become clear for all the relevant partners that an identifier, e.g., a famous author is useful for both libraries, archives and other cultural heritage institutions (Malmsten 2013).

The next major phase started in September 2011 when the National Bibliography and Swedish Authority file, two subsets of the LIBRIS database, became available in the same format in which they had been created (MARC21). This decision was a strategic one in a sense to expose Resource Description Framework (RDF)/linked data derived from the records, and the records in their original form. By taking this step, anyone can evaluate reference and contribute to the work done by the Swedish National Library. Their assumption is that visibility and openness will lead to higher quality data. The data was published under the Creative Commons Zero license terms. The next main goal is to publish the whole catalogue dataset according to these license terms. The main challenge in this field to handle the non-LIBRIS origin records (comes from other libraries, data aggregators) according to the same license rules as the native ones (Malmsten 2013).

DATA.BNF.FR project in the French National Library

The data.bnf.fr project endeavours to make the data produced by Bibliothèque nationale de France (French National Library-BnF) more useful on the Web. It gathers various BnF resources and external resources on pages devoted to an author, a work, or a subject. These pages organize the Web contents, links, and services provided by BnF. Available online since July 2011, data.bnf.fr is still evolving and expanding (National Library of France 2014).

With data.bnf.fr, it is possible to:

- reach BnF resources directly from a Web page, without any previous knowledge of the services provided by the library;
- get oriented in the BnF resources and possibly find external resources.

The objective is to put forward the BnF’s collections and to provide a hub between different resources. Data.bnf.fr is meant to support the BnF’s other applications. The project belongs to the BnF’s policy of becoming part of the Web of data and adopting Semantic Web standards.

Data.bnf.fr and Gallica have won the Stanford Prize for Innovation in Research Libraries (SPIRL) (National Library of France 2014).

The main objectives are to:

- make the data produced by the BnF more visible on the Web;
- federate the data produced by the BnF, both within and outside the catalogues;
- contribute to collaboration and metadata exchange by creating links between structured and trustable resources;
- facilitate reuse of metadata (under Open License) by third parties. (ibid.)

The data model used in data.bnf.fr makes it possible to federate data extracted from internal applications, but also to include links to external sources. Resources produced by the BnF (authority and catalogue records, finding aids for archives and manuscripts, digital documents) are assigned permanent identifiers – ARK identifiers– that enable the creation of persistent links (ibid.).

The first step was to develop bibliographical frameworks that are tested at an international level, especially the „FRBR" model.

This step was followed by modelling efforts aiming at displaying the data in RDF (Resource Description Framework) on the Web of data. In the BnF's view, the implementation of these technical standards must ensure interoperability between external and internal databases, through machine readable and structured data (ibid.).

Exposing data in RDF

In the long term, useful, reliable and controlled data will be displayed and integrated in the growing world of the Web of data, by abiding to the semantic Web standards. This must be done in conformance with international initiatives to facilitate the use of informational or administrative public data (ibid.).

Being on the Web of data implies the use of specific technical solutions in order to create links: dereferenceable and permanent URIs (Uniform Resource Identifiers), a content negotiation mechanism, and an access to raw data (ibid.).

Linked Open Data fosters data exchange between library and other communities, and brings solutions for formats interoperability. The Deutsche Nationalbibliothek, the British Library, and the Library of Congress have also adopted these tools in order to open their bibliographic data (ibid.).

The reusable data that BnF display include subject authority records from the RAMEAU repository, which is used to index bibliographic records at the BnF. These records were converted to the RDF SKOS language (Simple Knowledge

Organization System), within the framework of the European project TELplus. This repository is now regularly updated on data.bnf.fr with inputs from the whole database maintained by the BnF (Ibid).

External links to data.bnf.fr

Data.bnf.fr is part of the Web and provides external links to Web sites, either maintained by the BnF or completely independent.

There are several kinds of links:

Links to other external repositories, to which data produced by the BnF is aligned, such as the Library of Congress, the Deutsche Nationalbibliothek, VIAF (Virtual International Authority File), IdRef, Geonames, Agrovoc, and Thesaurus W (the French National Archives' thesaurus).

Links to search forms in which query terms (author name, subject, work title) are automatically pre-typed: BnF catalogue général, CCFr, BnF archives et manuscripts, CNLJ-La Joie par les livres, Europeana, SUDOC (Système universitaire de documentation), Worldcat, Wikipedia.

Wikipedia provides thumbnails for authors, whenever no one could be found on Gallica, and a short biography. This data is retrieved through Dbpedia.

The data belongs to separate databases. It is produced and stored in different formats. Data.bnf.fr extracts, transforms and gathers datasets in a unique database and makes them interoperable (ibid.).

The working tools of the model:

- Unique and permanent identifiers assigned to every record: the BnF uses ARK identifiers for records from the Catalogue general and digital documents from Gallica,
- Bibliographical description standards;
- Authority records for persons, corporate bodies, works and subjects, and data matching and federation techniques;
- Rely on authority records, which form the basis for all author, work, and subject pages, in order to gather and organise the different data silos. The different resources are collocated through the authority record's identifier;
- Author pages collocate all bibliographic records that are linked to the author's identifier;
- Work pages collocate all records that are linked to both the author's and the work's identifier. When there is no link to the work authority record, there is a simple matching mechanism based on string recognition techniques („words beginning with"). „Subject" pages

collocate all records that have a link to the same subject (National Library of France 2014).

The Hungarian National Library: First national semantic web project

The National Széchényi Library (NSZL) published its entire OPAC and Digital Library and the corresponding authority data as Linked Open Data in 2010 as one of the first public collections in Europe. The used vocabularies are RDFDC (Dublin Core) – (MARCXML to RDF/XML conversion with XSLT for OPAC bibliographic data) FOAF for names, and SKOS for subject terms and geographical names. NSZL uses CoolURIs. Every resource has both RDF and HTML representations. The RDFDC, FOAF and SKOS statements are linked together. The name authority dataset is matched with the DBPedia (semantic version of Wikipedia) name files. NSZL also supports the HTML link auto-discovery service. The available linked data resources are the following: dataset of person names, subject authority dataset, catalogue records in semantic forms. These can be searched and retrieved to external resources in the semantic web via an SPARQL endpoint (Horváth 2011a).

There was no specific project related to this field in the library, although small developments pointed to the same direction. Three members of the directorate of informatics developed it when time permitted. In 2009 they realized that they had almost everything in order to publish linked data on the semantic web. They converted the library thesaurus to SKOS format. Via the LibriURI tools the OPAC records have become accessible via URL. The URL-based search in the NSZL integrated system has become available via the SRU protocol with the Yaz proxy tool. They could use the experiences of the Swedish National Library (LIBRIS) semantic web implementation project. The main aims were the following: Library datasets need to be open (get your data out), need to be linkable, and also need to provide links. Datasets must be part of the network, cannot be an end in itself and the system must allow hackability.

The major advantages of the RDF-based semantic model are the following:

- RDF Clients can look up every URI in the RDF graph over the Web to retrieve additional information;
- Information from different sources merge naturally;
- RDF links between data from different sources can be set;

Information expressed in different schema, can be represented in a single model (Horváth 2011b).

The model can be described simply in the following way. The content location and the representation ways of the manifestation of the content can be found in RDF and HTML formats:

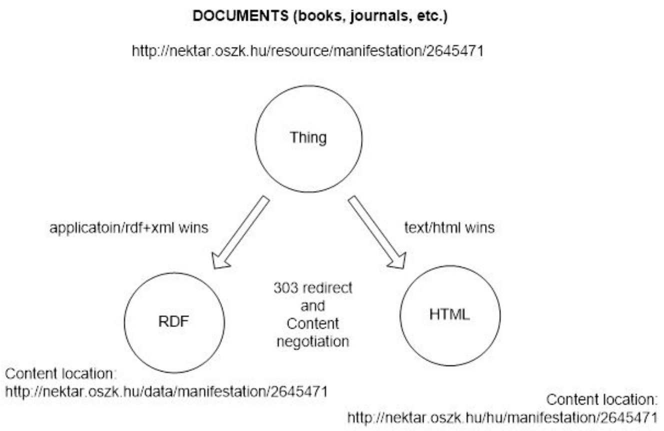
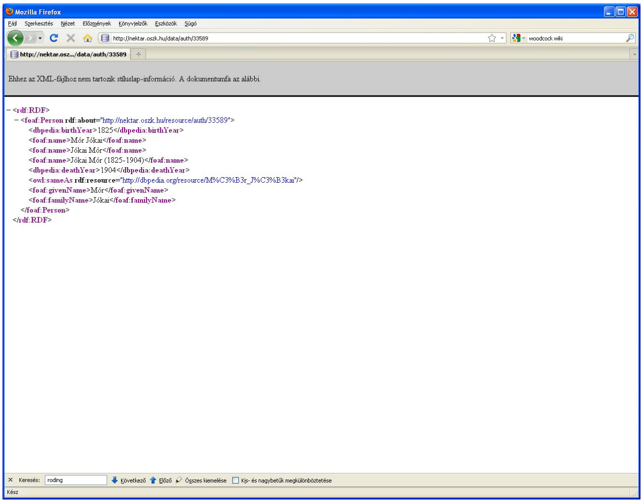

DOCUMENTS (books, journals, etc.)
http://nektar.oszk.hu/resource/manifestation/2645471
Thing
applicatoin/rdf+xml wins
text/html wins
303 redirect and Content negotiation
RDF
HTML
Content location: http://nektar.oszk.hu/data/manifestation/2645471
Content location: http://nektar.oszk.hu/hu/manifestation/2645471

Figure 1. An example of document display outputs (RDF and HTML) from the National Széchényi Library Catalogue (Horváth, 2011 b)

–If application/rdf+xml is accepted the xml is given from this address via content negotiation and 303 redirect: http://nektar.oszk.hu/data/manifestation/2645471

– If text/html is accepted: Depending on the language of the browser either the Hungarian or the English interface of the OPAC (LibriVision) is given. The default is Hungarian (again via content negotiation): http://nektar.oszk.hu/hu/manifestation/2645471)


Mozilla Firefox
Fájl Szerkesztés Nézet Előzmények Könyvjelzők Eszközök Súgó
http://nektar.oszk.hu/data/auth/33589
woodcock wiki
http://nektar.osz.../data/auth/33589
Keresés: roding
Következő Előző Összes kiemelése
□ Kis- és nagybetűk megkülönböztetése
Kész

Ehhez az XML-fájlhoz nem tartozik stíluslap-információ. A dokumentumfa az alábbi.

```
- <rdf:RDF>
  - <foaf:Person rdf:about="http://nektar.oszk.hu/resource/auth/33589">
      <dbpedia:birthYear>1825</dbpedia:birthYear>
      <foaf:name>Mór Jókai</foaf:name>
      <foaf:name>Jókai Mór</foaf:name>
      <foaf:name>Jókai Mór (1825-1904)</foaf:name>
      <dbpedia:deathYear>1904</dbpedia:deathYear>
      <owl:sameAs rdf:resource="http://dbpedia.org/resource/M%C3%B3r_J%C3%B3kai"/>
      <foaf:givenName>Mór</foaf:givenName>
      <foaf:familyName>Jókai</foaf:familyName>
    </foaf:Person>
  </rdf:RDF>
```

Figure 2. Different kind of namespace resources about the same author (Mór Jókai) integrated into an RDF record (Horváth 2010)

Different URI -s can be used for the same resource. For example, Mór Jókai Hungarian writer (born in Komárno, currently in Slovakia) can be identified with two URI-s http://nektar.oszk.hu/resource/auth/33589 in the NSZL library database and http:// http://dbpedia.org/resource/M dbpedia.org/resource/M in

DBPedia. The owl: SameAs links are resolving this problem. With this kind of link DBPedia can attach the VIAF link of the same person to its semantic interface and the different language versions of the same entry also (see at: http://dbpedia.org/page/M%C3%B3r_J%C3%B3kai) (Horváth 2010).

ALIADA project–a short overview

The international ALIADA project, financed by the EU, focused on creating tools for the implementation of semantic web to public collection environment. The project partners came from different European countries and different sectors (software developing industry, library, museums, and archives) (Horváth 2015). The starting point is easy in a sense that GLAM (Galleries, Museums, Archives and Libraries) institutions have usually rich metadata related to their collections. Metadata was mainly stored in standard schemes (MARC, Dublin Core, Lido etc.) Through the ALIADA framework, various metadata subsets from library, archives, gallery museum management systems (GLAM catalogue data, bibliographic data and authority data) can be converted from standard metadata input forms (e.g. MARC, LIDO, DUBLIN CORE) into RDF based semantic compatible format according to the ALIADA ontology. (Aliada Project 2015) The conversion process is made with an open source Java software. Data subsets are stored in a Virtuoso database and exported to a datadump file that is publicly available online. All the semantic data subsets through a SPARQL endpoint are registered in the datahub.io database with standard descriptions, links to the subsets and the address of the semantic Virtuoso database. Even before the automatic publication of semantic datasets in the semantic cloud, these can be linked to other datasets. The ALIADA software also automatizes the whole conversion and publication process. The partner institutions have to provide only standard metadata input subsets; the public collection experts do not need deep expertise on semantic web technologies. The semantic datasets can be linked to other datasets, such as Europeana, British National Bibliography, Spanish National Library, Freebase Visual Art, DBpedia, Hungarian National Library, Library of Congress Subject Headings, Lobid, MARC codes list, VIAF Virtual International Authority File or Open Library (Horváth 2014).

The project just finished in October 2015. The ALIADA software tool is free and publicly available to all the interested parties under the terms of the GNU GPL v3 (Aliada Project 2015, Horváth 2014).

An example of practical use of ALIADA framework tool will be described in the next section.

A case study of a semantic web related project in the Petőfi Literary Museum with an integrated library system in museum environment

An example of practical use of a semantic web based database is to build a triple store from a part of the database of the Qulto integrated library and museum automation system of the Petőfi Literary Museum (PIM, the abbreviation of its Hungarian name: Petőfi Irodalmi Múzeum). The museum's duty is to collect documents and objects connecting to the important personalities of the Hungarian literature. The museum's library also collects the documents of this theme, and the bibliographic descriptions of the library catalogue use the records, and the descriptive metadata of the museum inventory items. These catalogue items also contain important additional information about the novelists who are, as authors or mentioned personalities, included in the museum records (Bánki and Mészáros 2016).

The common information contained by a name authority record in a library catalogue or in a museum electronic inventory system, are personal name, date of birth and death, title, profession, data sources and linked bibliographic data, but there are a plenty of attributes in the catalogue of the library automation system of PIM, added to the records e.g. prices, exact date of birth and death, place of birth, death and residence, parents, husband, wife, children, sex, religion, education, jobs, important life events of the novelist etc. From these attributes a complex information package was prepared and stored in the Qulto integrated collection management system (ICMS) of the PIM. Most of the information was not imported in the Qulto ICMS, but in 22 separate Access databases, which were used by the experts of the museum for ten years before the Qulto system had been introduced. Data conversion was necessary from the Access based system to Qulto, and after the migration the information had been added to separate authority records. These information units had to be merged into one, main record from the various data items. In three steps more than 110,000 name authority records were selected as duplicated records. Duplication means that another name authority record was found in the database of PIM as a main name record describing the same person. After the information could be merged from the duplicated record to its main pair, the corrected database was ready to become one of the base data store elements of the Hungarian National Namespace. At the same time, it was also published on semantic web. So after consolidating the name database, and the record number was decreased to 620,000 items in the Qulto database of Petőfi Literary Museum, the dataset was ready to be uploaded somewhere or to be prepared as a local triple store.

There were three possibilities for us to publish the authority records of the collection Management System of Petőfi Literary museum on semantic web:

1. Load it to VIAF;

2. Build a triple store working together with other Hungarian museums, for example, Hungarian National Museum, or Museum of Fine Arts;
3. Create an own triple store in the PIM.

Each option will be described in the following sub-sections.

Load to VIAF

An option for publication is to connect to OCLC and load it to the VIAF database. As we have already mentioned above The VIAF–Virtual International Authority File–is, as an important unit of semantic web, coordinated by OCLC. It is based mostly on the authority records of libraries, so it works like a library catalogue in this sense. The identification of personalities is based mostly not on the metadata of the name records, the personal and biographic information of the novelists, but on the bibliographic data linked to them, the documents that were written by or about them. This way of identification is convenient for a library, having usually the name records of authors, but not for the factographic database of a museum. These institutions do not have many books, but have information entered from reference books. They have to identify the units of the authority database, the persons themselves, by the attributes of the biographic data.

Otherwise the upload is useful, necessary, and hopefully VIAF can use the uploaded records. We first connected to OCLC and sent a dataset of the first trial version of authority data export file, containing authority records having already linked bibliographic data in the local library catalogue. The VIAF needed a MARC21 export, which should have been prepared, creating a HUNMARC-MARC21 conversion from the Qulto ICMS, which as a MARC based system uses the Hungarian national standard HUNMARC as an internal data storage format.

The VIAF has already got a plenty of name data from Hungary, hopefully these name elements will be automatically identified by the system of VIAF, and the existing records in the VIAF database will be enriched by the newly sent data, and also new records will be created from the personal database sent from PIM to VIAF. Therefore, creating an authority export for VIAF upload, the personal names which had bibliographic records in the database and had enough additional information as authority records were selected.

Using the ALIADA application that was already installed in Hungary

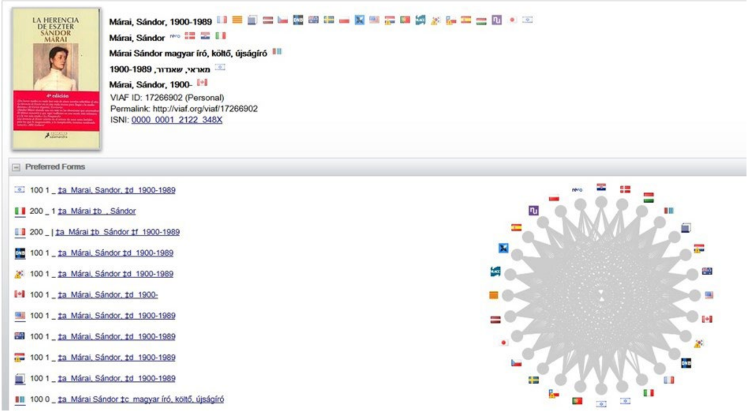
We have already provided a short overview about the ALIADA application (software tool framework) in the previous section. Here we are offering a practical example of its use.

The Museum of Fine Arts of Budapest has built its own Aliada database, with the possibility to define more sub-users and sub-databases. (http://www.szepmuveszeti.hu/aliada_en). The museum has published the descriptions of its 4000 artefacts on the semantic web with the help of ALIADA tool, and also gave the possibility to the Petőfi Literary museum to try this application, both the input and the web based public interface.

The workflow of data upload by the ALIADA pilot project was the same as by the OCLC. First we had to choose the records to be uploaded. The aspects of selection were almost the same, so the records had to contain enough information, they had to be entered into the proper sub-databases. It is possible to upload to ALIADA those authority records which do not have any bibliographic records joined to them. Thanks to the six year joint effort of the experts of the museum and library and the Qulto software support, the redundancy of the database was almost fully decreased. On the other hand, it was necessary to control and filter the duplicates of the uploaded names from the database in ALIADA. The existence of obligatory data elements had to be checked also. As in the case of VIAF export some data manipulation was necessary, e.g. the bibliographic data links were filled also into the authority data, to make it the integral part of the MARC authority record.

The Qulto internal format based on the structure of MARC, and it has its own structural logic, so the authority data have their quasi authority elements. For example, an authority record can be joined to corporate or geographical name records in a hypertext seeming data network. All these attached sub authority elements had to be appended to the authority output, and a proper MARC 21 header had to be prepared by the MARC authority export as well. In the past 6 years the PIM personal authority records were developed to be able to contain plenty of various information, in various MARC fields and subfields, not defined in the default MARC21 standard. These new data elements had to be mapped to the MARC21 data fields, recognizable for ALIADA MARC21 import format. In the future we'll try to enhance the acceptable field list of ALIADA MARC import. During the MARC import ALIADA converts the authority MARC data to RDF statements. ALIADA is a user-friendly and easy to use application. The operator has to validate the input data set, has to select the sub-database (graph), and delete the unnecessary records from the Virtuoso database. The ALIADA import program always adds elements, but never merge duplicated records. You have to filter your dataset from duplicated records before the ALIADA import. If necessary, the demanded data type can be selected, data fields and subfields can be marked in order to be converted through the import process. There is a problem by import in the pilot project: a relatively small size of input files was accepted by ALIADA.

The result is the converted dataset, into the Virtuoso database, which is browesable, containing valid data links generated by ALIADA. The dataset can be inserted to the semantic cloud. Another possibility is to join data elements automatically with other ones, and these links can be added to the local database to enrich authority or bibliographic records with other data connections. Also the VIAF URI-s can be added to the authority records in the local database.



Márai, Sándor, 1900-1989

Márai, Sándor

Márai Sándor magyar író, költő, újságíró

מאראי, שאנדור, 1900-1989

Márai, Sándor, 1900-

VIAF ID: 17266902 (Personal)
Permalink: http://viaf.org/viaf/17266902
ISNI: 0000 0001 2122 348X

Preferred Forms

100 1 _ ‡a Marai, Sandor, ‡d 1900-1989
200 _ 1 ‡a Márai ‡b , Sándor
200 _ | ‡a Márai ‡b Sándor ‡f 1900-1989
100 1 _ ‡a Márai, Sándor ‡d 1900-1989
100 1 _ ‡a Márai, Sándor ‡d 1900-1989
100 1 _ ‡a Márai, Sándor, ‡d 1900-
100 1 _ ‡a Márai, Sándor, ‡d 1900-1989
100 1 _ ‡a Márai, Sándor, ‡d 1900-1989
100 1 _ ‡a Márai, Sándor, ‡d 1900-1989
100 1 _ ‡a Márai, Sándor, ‡d 1900-1989
100 0 _ ‡a Márai Sándor ‡c magyar író, költő, újságíró

Figure 3. Example of a VIAF record

Our goal is to further enhance this semi- automatized workflow (that has built from these four steps: 1. data manipulation in PIM Qulto database 2. data conversion from HUNMARC to MARC21 3. preparing MARC XML from relational database quasi MARC data units 4. ALIADA import converting to RDF statements of Virtuoso database) to develop a fully automatized one. Also the VIAF upload is planned to be automatized.

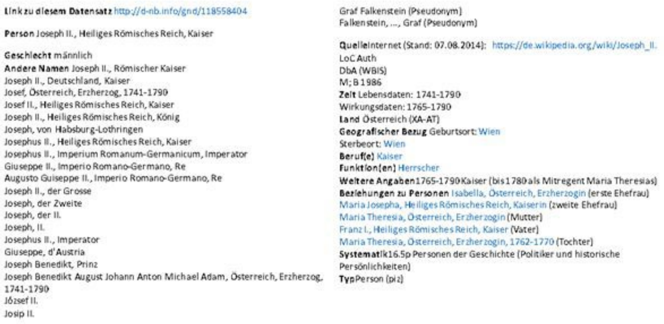
An own triple-store in PIM

The third possibility, to build an own triple store of PIM, potentially means to install an own Aliada application to the local server of Petőfi Literary Museum. After the RDF statements were controlled, the new database is ready–several output formats to be prepared from it: FRBRoo, WGS84, SKOS, SKOSXL, FOAF, DCTERM, and OWL-TIME.

The advantages of Qulto as a Library or Museum Collection management software is the almost unlimited possibility of defining new special data elements, and also the highly customized segmentation of the records. So all the various information segments can be added in separate and specially marked record fields. In this way any type of outputs and export data sets can be produced from this input of records.

The potential aim of use of semantic web databases and database elements, is to identify and describe persons who are hardly definable by name strings, or have many connections and are enriched with plenty of sub-elements.

An example below is of the famous Habsburg emperor Joseph II, having a well-known and often mentioned but very short name, which is hard to identify by entering search terms, and has a long name with plenty of titles and Christian names. The record (for he was emperor of the Holy Roman Empire, i.e. the ruler of Germany of that time) for Deutsche Bibliothek tries to identify him as shown in Figure 4.



Link zu diesem Datensatz http://d-nb.info/gnd/118558404

Person Joseph II., Heiliges Römisches Reich, Kaiser

Geschlecht männlich
Andere Namen Joseph II., Römischer Kaiser
Joseph II., Deutschland, Kaiser
Josef, Österreich, Erzherzog, 1741-1790
Josef II., Heiliges Römisches Reich, Kaiser
Joseph II., Heiliges Römisches Reich, König
Joseph, von Habsburg-Lothringen
Josephus II., Heiliges Römisches Reich, Kaiser
Josephus II., Imperium Romanum-Germanicum, Imperator
Giuseppe II., Imperio Romano-Germano, Re
Augusto Guiseppe II., Imperio Romano-Germano, Re
Joseph II., der Grosse
Joseph, der Zweite
Joseph, der II.
Joseph, II.
Josephus II., Imperator
Giuseppe, d'Austria
Joseph Benedikt, Prinz
Joseph Benedikt August Johann Anton Michael Adam, Österreich, Erzherzog, 1741-1790
József II.
Josip II.
Graf Falkenstein (Pseudonym)
Falkenstein, ..., Graf (Pseudonym)

QuelleInternet (Stand: 07.08.2014): https://de.wikipedia.org/wiki/Joseph_II.
LoC Auth
DbA (WBIS)
M; B 1986
Zeit Lebensdaten: 1741-1790
Wirkungsdaten: 1765-1790
Land Österreich (XA-AT)
Geografischer Bezug Geburtsort: Wien
Sterbeort: Wien
Beruf(e) Kaiser
Funktion(en) Herrscher
Weitere Angaben1765-1790Kaiser (bis 1780 als Mitregent Maria Theresias)
Beziehungen zu Personen Isabella, Österreich, Erzherzogin (erste Ehefrau)
Maria Josepha, Heiliges Römisches Reich, Kaiserin (zweite Ehefrau)
Maria Theresia, Österreich, Erzherzogin (Mutter)
Franz I., Heiliges Römisches Reich, Kaiser (Vater)
Maria Theresia, Österreich, Erzherzogin, 1762-1770 (Tochter)
Systematik16.5p Personen der Geschichte (Politiker und historische Persönlichkeiten)
TypPerson (piz)

Figure 4. Semantically enriched catalogue record from the Deutsche Bibliothek for Joseph II.

Future development plans: Museumap (integrated museum portal) and KÖZTAURUSZ (Universal thesaurus of the National Széchényi Library, public libraries, scientific and technical libraries)

Museumap is the central portal of Hungarian Museums managed by the Hungarian National Museum. This portal offers a central search function in an aggregated museum database of museum catalogues. Nomination and location data of digital inventory books from the different museums are available in a thesaurus (also in English). In the future it can be a basic dataset of a new museum namespace model.

The main aim of the KÖZTAURUSZ is to cover all possible subject fields, professions and terms that are available in library catalogues, i.e. it is really general. This thesaurus has already been integrated with most of the library integrated systems in Hungary. It is already compatible with SKOS. As an SKOS-based semantic dataset it can be an important general element of the central Hungarian namespace in the future. The concept of this central namespace will be described in the following section.

The concept of the Hungarian National Namespace

The Hungarian National Digital Archive, National Széchényi Library, Hungarian National Archive, and Petőfi Literary museum started a broad collaboration project in 2012 with the national namespace concept. The idea of the concept is to focus on the experiences of the process of building large collaborative namespaces as professional service frameworks created by international collaboration (GETTY, VIAF, ICONCLASS). It has become clear for the members of the project that the focus of future professional activities will be on the usage of namespaces in the semantic web environment. The target groups are both the institutions with their professional needs and the users of the public collections in the broadest sense. Professionals from different institutions agree that the institutional namespaces and other high-quality semantic-web compatible metadata collections (such as KÖZTAURUSZ that was mentioned above) will form the basis of the common Hungarian namespace. Following this concept, cleaning the dataset of names, including the identification and data clearance of entities, has been a major contribution by the Petőfi Literary Museum towards the creation of the Hungarian national namespace.

The creation process of the national namespace is still in progress. However, the concept is already available. The collections from the different institutions can be converted into semantic-web compatible format. In this way these can be fully connected and are interoperable wi th each other. Using joint namespaces seems to be the best way to represent the Hungarian cultural heritage as a whole, describing all of the available contexts and connections through the different segments of this broad area. All the cultural databases could be available through a common interface regardless of the traditional institutional barriers among the archive, library and museum sectors. Collections and datasets being stored by the individual institutions can be represented in their own local contexts. Furthermore, the public can have access to the digital cultural heritage as a whole. The content of the local collections could be represented through a framework of an integrated system (based on namespaces with common standards on the semantic web). The primary user group of the future national namespace will be the general public. Through the first step of the implementation process, however, the collaborating institutions must focus on their own institutional contexts and needs in order to create their own high quality datasets, service workflows, and infrastructure. The second step in the future should focus on the implementation of all the services and functionalities focusing on the general public. Another strategic goal of the Hungarian National namespace can be a data authority function. The National Namespace must serve qualified, controlled and clear data on request to anyone, either to a public collection and other public institutions or to any individuals.

Schema.org and microdata: New semantic web tools in the HTML5 standard

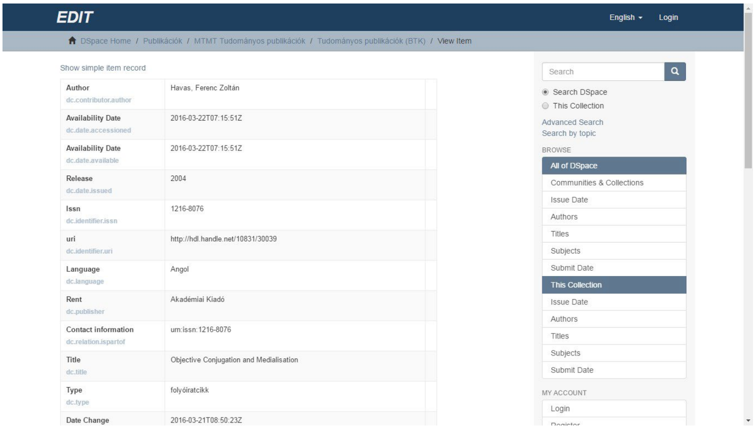
Many librarians are familiar with basics of the HTML language. Usually, HTML tags tell the browser how to display the information included in the tag. For example, <h1>Avatar</h1> tells the browser to display the text string „Avatar" in a heading 1 format. However, the HTML tag doesn't give any information about what that text string means–"Avatar" could refer to the hugely successful

3D movie, or it could refer to a type of profile picture–and this can make it more difficult for search engines to intelligently display relevant content to a user. The web of documents is linking documents links which are not qualified. Otherwise on the semantic web we are linking datasets with qualified links. Schema.org simply provides a collection of shared vocabularies that can be used to mark up the public collection homepages (and any other homepages of course) in ways that can be understood by the major search engines: Google, Microsoft, Yandex and Yahoo. You can use the schema.org vocabulary along with the Microdata, RDFa, or JSON-LD formats to add information to your Web content ("Getting Started with Schema.org Using Microdata" 2016). In case of RDFa, the RDF statements are properties of HTML tags and can be generated as a collection of HTML-based homepage texts.

Why are microdata and microformats useful? The web pages have an underlying meaning that people understand when they read them. But search engines have a limited understanding of what is being discussed on those pages. By adding additional semantic tags (for example with RDFa format) to the HTML of your web pages—tags that say, „Hey search engine, this information describes this specific movie, or place, or person, or video"—you can help search engines and other applications better understand your content and display it in a useful, relevant way. Microdata is a set of tags, introduced with HTML5, that allows you to do this (Horváth 2016).

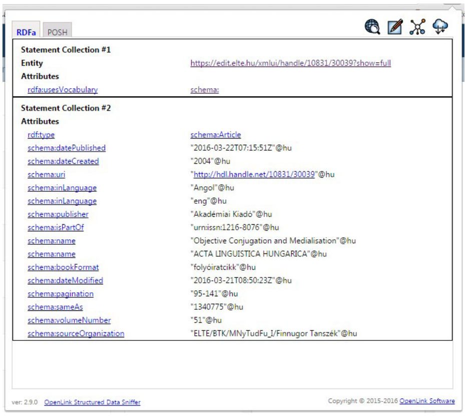
In Libraries with the help of Schema.org you can use the Library class and define FRBR-like attributes on the homepages (exampleOfWork, workExample). It is possible to define also connections (hasPart, isPartOf). Currently, microformats (schema.org and RDFa) are being used in OPAC (WorldCat, Koha), and in discovery systems (like VuFind) and repositories (like DSpace).

Here are some examples:



Author dc.contributor.author	Havas, Ferenc Zoltán
Availability Date dc.date.accessioned	2016-03-22T07:15:51Z
Availability Date dc.date.available	2016-03-22T07:15:51Z
Release dc.date.issued	2004
Issn dc.identifier.issn	1216-8076
uri dc.identifier.uri	http://hdl.handle.net/10831/30039
Language dc.language	Angol
Rent dc.publisher	Akadémiai Kiadó
Contact information dc.relation.ispartof	urn:issn:1216-8076
Title dc.title	Objective Conjugation and Medialisation
Type dc.type	folyóiratcikk
Date Change	2016-03-21T08:50:23Z

Figure 5. A sample record with semantic microformat tags in EDIT repository



Statement Collection #1	
Entity	https://edit.elte.hu/xmlui/handle/10831/30039?show=full
Attributes	
rdfa:usesVocabulary	schema:

Statement Collection #2	
Attributes	
rdf:type	schema:Article
schema:datePublished	"2016-03-22T07:15:51Z"@hu
schema:dateCreated	"2004"@hu
schema:uri	"http://hdl.handle.net/10831/30039"@hu
schema:inLanguage	"Angol"@hu
schema:inLanguage	"eng"@hu
schema:publisher	"Akadémiai Kiadó"@hu
schema:isPartOf	"urn:issn:1216-8076"@hu
schema:name	"Objective Conjugation and Medialisation"@hu
schema:name	"ACTA LINGUISTICA HUNGARICA"@hu
schema:bookFormat	"folyóiratcikk"@hu
schema:dateModified	"2016-03-21T08:50:23Z"@hu
schema:pagination	"95-141"@hu
schema:sameAs	"1340775"@hu
schema:volumeNumber	"51"@hu
schema:sourceOrganization	"ELTE/BTK/MNyTudFu_I/Finnugor Tanszék"@hu

ver: 2.9.0 OpenLink Structured Data Sniffer Copyright © 2015-2016 OpenLink Software

Figure 6. Sample of semantic statements in schema.org and RDFa (Horváth 2016)

In Hungary the first implementation of microformat tags can be found in the university library of the most traditional university in Budapest, Eötvös Loránd University (ELTE). The pages of the DSPACE-based institutional repository: ELTE Digital Institutional Repository (EDIT) are tagged with RDFa and Schema.org tags. Microformats will be used soon also in the VuFind based new integrated portal of the Hungarian National Library (support of microformats is a built-in function of VuFind) (Horváth 2016). Implementing microformats into online full-text databases in libraries can be a major step forward also in order to offer more semantic web-compatible data by these institutions with a relatively low level of efforts.

Conclusion

In this conclusion the results that the public collections have already achieved in semantic web field will be summarized and some challenges for the future will be highlighted.

It has already become clear that library and museum catalogues are natural inputs of databases being published as linked open data. In order to make a successful semantic conversion we have to get authority data from the catalogues. In the next step we have to de-duplicate and validate the authority record sets. The management of authority record can be successfully managed within a National Namespace Network in the future in Hungary.

Based on the validated authority data triple-stores, namespaces and ontologies can be built-up in order to publish authority data as Linked Open Data on the semantic web. Recently, the public collections mainly focus on building up their own semantic datasets and share them with other institutions on the semantic cloud.

The next step can be to provide user-friendly service models by offering highly valuable semantic content in an attractive user-friendly environment.

Followed by the implementation of these kind of systems in the future, the general public can give valuable feedback in order to correct the content of the various datasets and also enrich them with additional information.

By taking a look at the technical background of semantic data management, a major question is if the existing library management system (LMS) environment will be sufficient to manage the new (and constantly developing) semantic data standard requirements. It means that instead of the MARC data exchange format new data exchange standards can be implemented (BIBFRAME, RDA compatibility etc.). New social network compatible functions also can be applied to LMS User display interfaces. The other way can be that we perhaps need new types of tools or system architectures for this purpose in the future beyond the LMS system implementations. The public collection environments will be strongly integrated with each other in the digital sphere. However, the different historical traditions, professional practices and attitudes of museum staff, librarians, and archivist's can be handled as a challenge as well.

The representation of cultural heritage can turn into a new dimension but we can handle many kinds of challenges. A common responsibility of the cultural policy makers and cultural heritage professionals is to offer new and even more comprehensive ways to access our common cultural heritage.

References

ALIADA Project. 2015. "ALIADA as an open source solution to easily published linked data for libraries and museums." http://www.slideshare.net/aliadaproject/swib15-aliada.

Aliada Project. 2015. "Introduction to ALIADA Webinar." http://www.slideshare.net/aliadaproject/introduction-to-aliada-webinar.

Bánki, Zsolt, and Tibor Mészáros. 2016. "Checking the identity of entities by machine algorythms is the next step to the national name authorities." https://conference.niif.hu/event/5/session/14/contribution/26.

Bánki, Zsolt, Tibor Mészáros, Márton Németh, András Simon. 2016. "Checking the identity of entities by machine algorithms: The next step to the Hungarian National Namespace." Code4Lib Journal 33: 1-8. http://journal.code4lib.org/articles/11765

"Getting Started with Schema.org Using Microdata." 2016. http://schema.org/docs/gs.html.

Horváth, Ádám. 2010. "Linked Data at the National Széchényi Library : Road to the publication." http://swib.org/swib10/vortraege/swib10_horvath.ppt.

Horváth, Ádám. 2011a. "National Széchényi Library semantic web wiki." http://nektar.oszk.hu/wiki/Semantic_web.

Horváth, Ádám. 2011b. "Linked Data at NSZL." http://nektar.oszk.hu/w/images/0/04/LinkedDataAtNszl_06.pdf.

Horváth, Ádám. 2014. "The European ALIADA project." In Rome: 33rd ADLUG Annual Meeting, 1–20. http://www.slideshare.net/aliadaproject/aliada-intro-adamhorvath03.

Horváth, Ádám. 2016. "RDFa - Schema.org: Unity of document and semantic web." https://conference.niif.hu/event/5/session/10/contribution/27/material/slides/0.ppt.

Malmsten, Martin. 2013. "Cataloguing in the open: The disintegration and distribution of the record." Italian Journal of Library %26 Information Science 4, 1: 417–23. 10.4403/jlis.it-5512.

National Library of France. 2014. "About Data.bnf.fr." http://data.bnf.fr/about.

Németh, Márton, and András Simon. 2016. "Public collections on the semantic web in a Hungarian context". ITLib 2016, 2: 66-74