ZBORNÍK FILOZOFICKEJ FAKULTY UNIVERZITY KOMENSKÉHO
Ročník XXVII KNIŽNIČNÁ A INFORMAČNÁ VEDA Bratislava 2017

WIDENING THE LIMITS OF COGNITIVE RECEPTION WITH ONLINE GRAPH DATABASES ON THE SEMANTIC WEB

Márton Németh

Introduction

The implementation of semantic web paradigms to public collection environment (archives, libraries, museums) leads to a paradigmshift in the fields of information search and retrieval and digital document management. It can be stated that the world wide web itself can be transformed to a global content management system. The basic point is that any kind of data comes from people's mind; digital or offline documents can be linked. Linked data appear as an approach and set of technical tools rather than a properly defined technical standard (Meehan 2014). The benefits of the web focus on data not just on documents. Any kind of data sets can be described in a standard way and can be linked with each other. RDF gives the basic shape to linked data. It is a standard model of data interchange on the world wide web. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all data consumers to be changed. RDF extends the linking structure of the Web to use URIs to name the relationship between things as well as the two ends of the link (it is referred to as a "triple"). Using this simple model, it allows structured and semi-structured data to be mixed, exposed, and shared across different applications (Resource Description Framework 2014). Another really important point is the legal status of the different linked datasets. Creative Commons and other license solutions offer cultural institutions the legal way to share their data as open data. Different kind of license models belong to data and content derived from data. It is really important that cultural heritage data must be shared without

any restrictions. On the other hand based on linked open data several kinds of business models can be built-up towards the end-users. These value-added services based on liked open data sources can be offered through different kind of legal and business frameworks. External partners can also be involved in this way.

The conversion of different collections from archives libraries and museums into semantic web compatible datasets means that these collections are appearing in an online graph database environment. Public collections have to build up semantic ontologies. Semantic ontology is an explicit specification of conceptualization. RDF/OWL language is appears as a representation way of ontologies. Datasets have an RDF/XML description format and are organized according standard namespace schemas. Namespaces can identify various types of data inputs in semantic environment (like thesauri, authority data, datasets of personal name collections etc.). Different datasets must be published as linked open data in order to build-up standard connections with other standard RDF/XML based datasets. Connection can be maintained by standard SPARQL endpoints. SPARQL is a document retrieval language. It is optimized to machine-to-machine communication. With that language it is possible to make queries on online graph databases that store linked open data from public collections. From the cognitive point of view it can be interesting that end-users can get information via semantic web compatible cataloguing tools and via machine-to machine communication but the association links the machines can retrieve from the graph have been built by people. New information retrieval forms can raise the effectiveness of research processes. A special field of digital curating focuses on how different kinds of ontologies and semantic datasets are compatible with each other in order to build up complete digital linked open data based ecosystems for the benefit of users.

In this article I would like to offer examples of combination and effective representation of the ways of complex information from different datasets. Semantic tools can in a way help us bypass language barriers. Challenges can also be managed by the usage of different terminologies in certain research contexts. On the technical background of these services the online graph databases appear. The databases based on properly cleaned linked data have some unique functions that can help broaden the limit of cognitive reception of cultural heritage information. To establish and adequately link datasets with each other, however is only the first step. The second step is the creation of different kinds of resource models by resources not well served by current standards. The third step is to develop applications, user interfaces as outcomes based on linked open datasets to represent different segments for the benefits of users. It is possible to combine results from different collections from all over the world to get a more detailed overview of a chosen segment of our cultural

heritage. It leads to greater visibility of the world wide web (including deep web resources) and better user experience can also be offered this way.

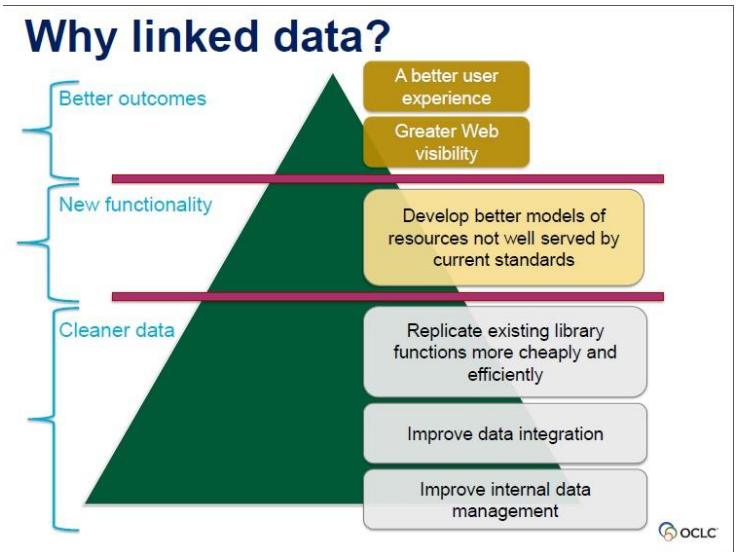


Fig. 1

Different levels of linked data management. This paper focuses on the better outcomes segment (Godby 2017).

1 Current implementation projects of Linked open datasets for public benefit

The data.bnf.fr project endeavours to make the data produced by the Bibliothèque Nationale de France (French National Library) more useful on the Web (National Library of France 2014). It retrieves and connects 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, the data.bnf.fr is still evolving and expanding.

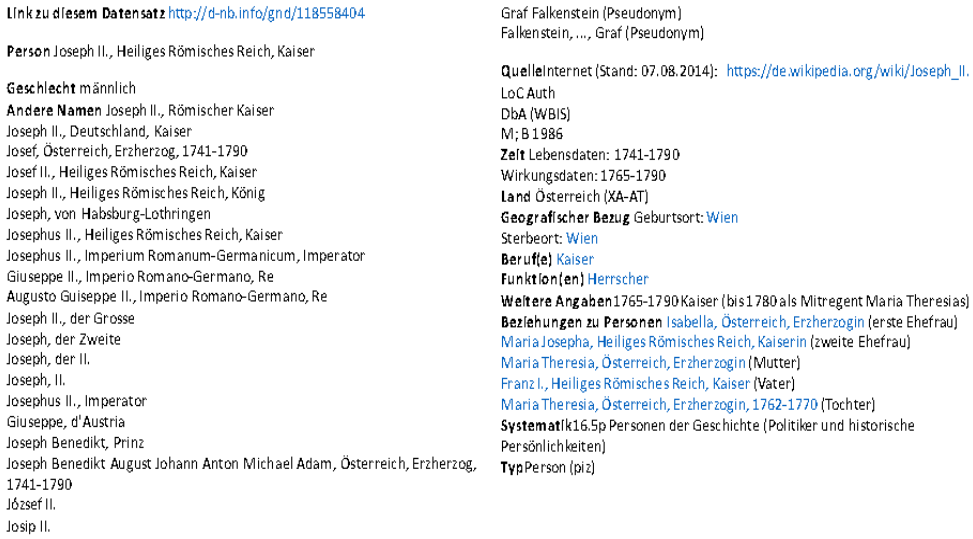
With data.bnf.fr, it is possible to:

- reach BnF resources directly from a Web page, without 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. The 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.

The model can be really relevant in our opinion because it can help to retrieve new ways of connecting different information resources devoted to a certain subject. A concrete example is one of the semantic representations of Victor Hugo's famous novel, Les Miserables (Victor Hugo Les Miserables 2016). Starting from this web page, 361 different editions of the novel can be found and compared, including digital documents. Different kinds of representations of the work (music recordings, theatre performance recordings) can be retrieved as well. Bibliography of corresponding literary articles and other documents can also be found. Other works and references also appear, including library catalogue links, pictures, diagrams or maps. The data of 121 contributors related to different editions and representations of this novel (literary people, e.g. illustrators, editors, translators, poets, playwrites, musicians, directors, filmmakers that have been inspired by the novel) are also available. The corresponding virtual exhibition materials and multimedia resources from the collections of the French National Library and other virtual collections can also be discovered. You can find direct link to the Wikipedia article of the novel, and you can search also on external websites (like Europeana) for corresponding information about this novel.

Another example of the power of sematic search is that of the Deutsche Bibliotek catalogue. It identifies all the corresponding major data sources on the Habsburg emperor Joseph II. The emperor has a well-known, frequently mentioned, but very short name which is hard to identify by entering search terms, while on the other hand he has a long name with plenty of titles and Christian names. In this example, important personal information and corresponding historical dates can be discovered. All the official alternative name forms in German and in all the major languages used in the former Habsburg countries also appear. The family relations are described as are the corresponding geographical locations, general titles and functions. The link of the corresponding Wikipedia article is also available. These data are linked in a graph, so we can search through any single set of information available and thus learn about the emperor. The multilingual description of the name is rather important. When searching for any name form in the semantic catalogue one can find all the information about the corresponding emperor.



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)

Quelle Internet (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 Angaben 1765-1790 Kaiser (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)
Systematik 16.5p Personen der Geschichte (Politiker und historische Persönlichkeiten)
Typ Person (piz)

Fig. 2

Example of a semantic description from the catalogue of Deutsche Bibliotek (Available from: http://dnb.info/gnd/118558404)

The third example is the model of Europeana. The cornerstone of the strategy of Europeana is to build an entity collection. It can be a service that can serve as a centralized point of reference and access to data on contextual entities (Manguinhas et al. 2016). Moreover, the main goal is caching and curating the data from the linked open data cloud. It can be a sort of Europeana knowledge graph. As the interlinking of data improves, it brings more context to objects, alleviates polysemy issues, expands language coverage, and helps build up a knowledge graph (web of data) that can be used by the third parties to improve their user experience (Manguinhas et al. 2016). Target vocabularies are used for semantic enrichment: Places - a subset of Geonames, corresponding to places which are part of the European countries and of some specific feature classes (214.307 resources). Agents - a subset of DBpedia, corresponding to most of the instances of dbp: Artists - with some exceptions, and integrated from 49 DBpedia language editions (165008 resources). Concepts - a subset of DBpedia, corresponding to a handful of concepts matching the needs from Europeana Collections (274 recources). Time Spans - the chronological periods from SemiumTime (2566 recources). Finding new dictionaries and keeping up-to-date information from external sources are essential as is the support of manual curating of both the existing and new entities. The main challenge to Europeana is the lack of core-referencing information to external disctionaries from data providers. Labels and values are not always accurate.

Sometimes relevant information, e.g. roles and professions, is missing. In some cases the coverage is needed to be expanded to other types of entities (like work and events) (Manguinhas et al. 2016).

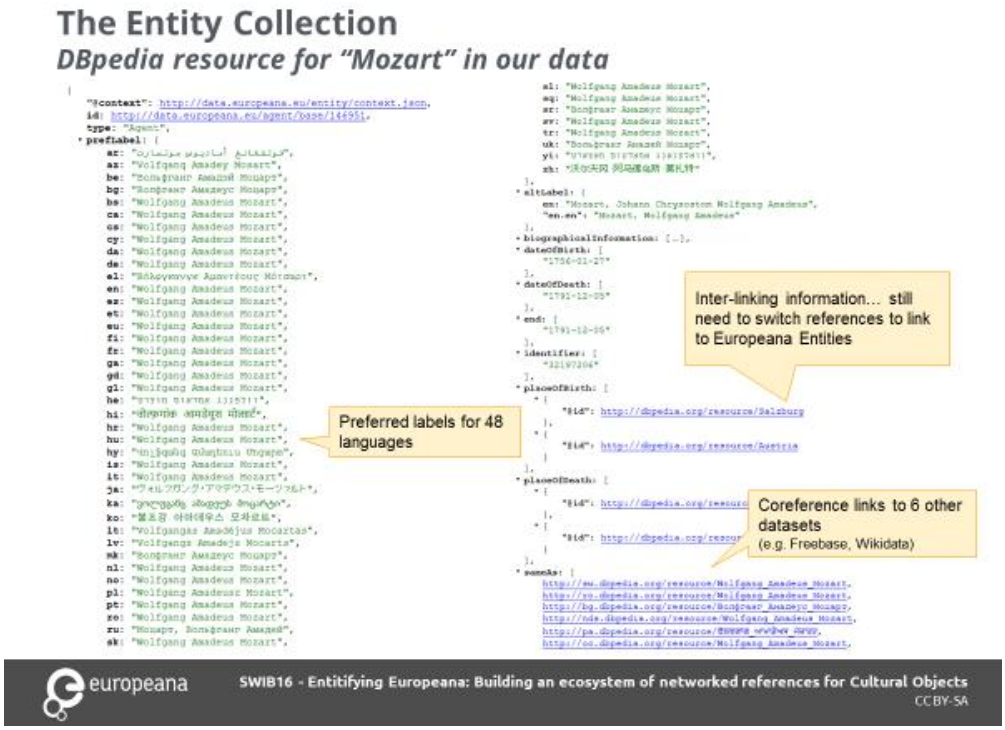


Fig. 3

Semantic enrichment form DBpedia to Europeana by resources to "Mozart" (Manguinhas et al. 2016)

The fourth set of examples belongs to the British Broadcasting Corporation (BBC) and its cooperation with different stakeholders. The BBC was among the first public media providers that started to build-up a complete linked open data portfolio. The BBC ontologies are built according to business requirements. They are all expected to evolve as do their requirements. The BBC produces a plethora of rich and diverse content about the things that matter to their audiences. The Linked Data gives them an opportunity to connect content and topics. They use ontologies to describe the world around us, the content the BBC creates, and the management, storage and sharing of the data within the Linked Data Platform.Things are the stuff that they talk about; they are the things that matter to the BBC's audience, and what the BBC makes content about. This set of data covers everything from athletes, politicians and music artists to football teams, subjects in the UK curricula and

places. They use domain ontologies to model the Things. They developed about a dozen of ontologies for modelling not just the different kind of news resources but the wildlife, food, and national education curricula within the United Kingdom (BBC 2016a). People, places and organisations that can be relevant to many domains are modelled in the CoreConcepts Ontology. As the BBC expands its usage of Linked Data, more of these domain-specific ontologies will evolve accordingly. They also have ontologies that support storage, management and ownership of Linked Data. The Provenance Ontology enables data to be stored in the appropriate named graphs with the necessary ownership and auditing information attached to them. The CMS Ontology allows for data management between systems and synchronization, between the Linked Data Platform and the content management systems where the content and Things are produced.

BBC has run several kinds of experimental projects based on linked open data resources since 2012 to empower innovative News storytelling across journalist workflows and audience experiences (BBC 2016c). One of them is the Juicer application interface that takes articles from the BBC and other news sites, automatically parses them and (based on their content) tags them with related DBpedia entities. The entities are grouped in four categories: People, Places, Organisations and Things (everything that does not fall in the first three ones) (BBC News Lab 2016a). The WAT webapp uses the Juicer API to generate a series of amusing graphs about who is talking about what across hundreds of news sources (BBC News Lab 2016b).

Another example of using the BBC Linked Open data resources is that of use of their resources to help improve the MusicBrainz database. When errors were found, the BBC fixed the mistake in the external data source and not within the walled garden of BBC's ICT infrastructure where only the BBC could benefit from the organisation's editorial expertise. Certainly, the BBC's charter requires the organisation to provide 'benefit' to the public, and contributing to the free and open MusicBrainz database fits nicely with public service remit. This is strategically a really smart approach and one of those quietly revolutionary ideas behind 'the Web as CMS'. The BBC's contributions add value to a resource, the MusicBrainz database. That added value, in turn, makes the resource more attractive to others who will use the resource and further improve the data (McGinnis 2016).

Another award-winning project of the BBC (Semantics Conference Linked Data Award, Leipzig, 2016) is the Research and Education Space (RES) (Semantics conference 2016). It aims to improve access to the UK's public archives for use in UK education and research, by facilitating the use of audio-visual and other archive media in teaching and learning. The idea of the RES project grew out of a desire to enrich the materials available for teaching and studying in the UK. It has been

developed as part of the BBC's Archive Strategy, to deliver significant public value from the BBC's own huge archive, by making as much of it as possible available to those in formal UK learning, from primary schools through to universities. The project comprises a partnership between the BBC, Learning on Screen and Jisc, working in collaboration with the GLAM and Education Technology sectors. An open platform (Acropolis) has been built by the BBC which organises and indexes the catalogues of public archives. A public service has run for the UK education - working with public sector organisations, e.g. The British Museum, the British Library, the Europeana, the BBC to release their digital collections as linked open data. The aim is for RES to make study, research, lessons and learning more interesting, varied, colourful and informative, to enrich teaching across different levels and subjects and to support research at all levels. It offers a reliable, openly licensed index of cultural assets from some of the world's leading cultural institutions. There is no charge to build on the RES platform. Video, audio, photos and images, commercial music, sheet music and historical documents will be available at all educational levels. The platform itself is up and running and the catalogue is growing, and a range of tools is being developed to be used by software companies and educational publishers who want to build resources around it. (BBC 2016b). The platform already includes 47,000 images from the BBC library, 1,650 classroom clips from the BBC Teach, 1,500 media items from the BBC RemArc (a Reminiscence Archive developed to benefit dementia patients and their carers) 10,000 BBC Playable programmes, including the BBC World Service content. This is in addition to the library of over 1 million BBC TV and Radio programmes that can be permanently served to educational establishments. The BBC Shakespeare Archive Resource launched at the end of 2015 to mark 400 years since Shakespeare's death, provides schools, colleges and universities across the UK with access to hundreds of BBC television and radio broadcasts of Shakespeare's plays, sonnets and documentaries about Shakespeare. Subtitles are being added to all the television programmes which means students, teachers and academics can now watch Shakespeare's plays and sonnets with their corresponding transcript. As much of the content pre-dates the use of subtitles, we are really excited to be offering users the chance to watch many BBC adaptations dating from the 1950s with subtitles for the very first time. The appearance of public collection data on the RES platform is expected from 2017 (Bishop 2016).

2 Schema.org and microdata: New semantic web tools of the HTML5 standard in digital library environment

HTML language is the basic markup language of the world wide web. Usually, HTML tags tell the browser how to display the information included in the tag. For example, <h1>Avatar</h1> tells the web browser to display the text string "Avatar" in a heading 1 format. However, the HTML tag does not give any information about what the 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 the relevant content to a user. The web of documents is linking documents links which are not qualified. On the semantic web we link datasets with qualified links. Schema.org simply provides a collection of shared vocabularies that can be used to mark the public collection homepages (and any other homepages) in ways that can be understood by the major search engines: Google, Microsoft, Yandex and Yahoo!. The schema.org vocabulary can be used along with the Microdata, RDFa, or JSON-LD formats to add information to your Web content (Getting started 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 to 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 digital libraries, one can use the Library class and define FRBR-like attributes on the homepages (exampleOfWork, workExample) with the help of Schema.org. It is also possible to define the connections (hasPart, isPartOf). Currently microformats (schema.org and RDFa) are being used in the OPAC (WorldCat, Koha), in discovery systems (like VuFind), and repositories (like DSpace).

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) has been tagged with the RDFa and Schema.org tags. Microformats will be used soon in the open source 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 effort.

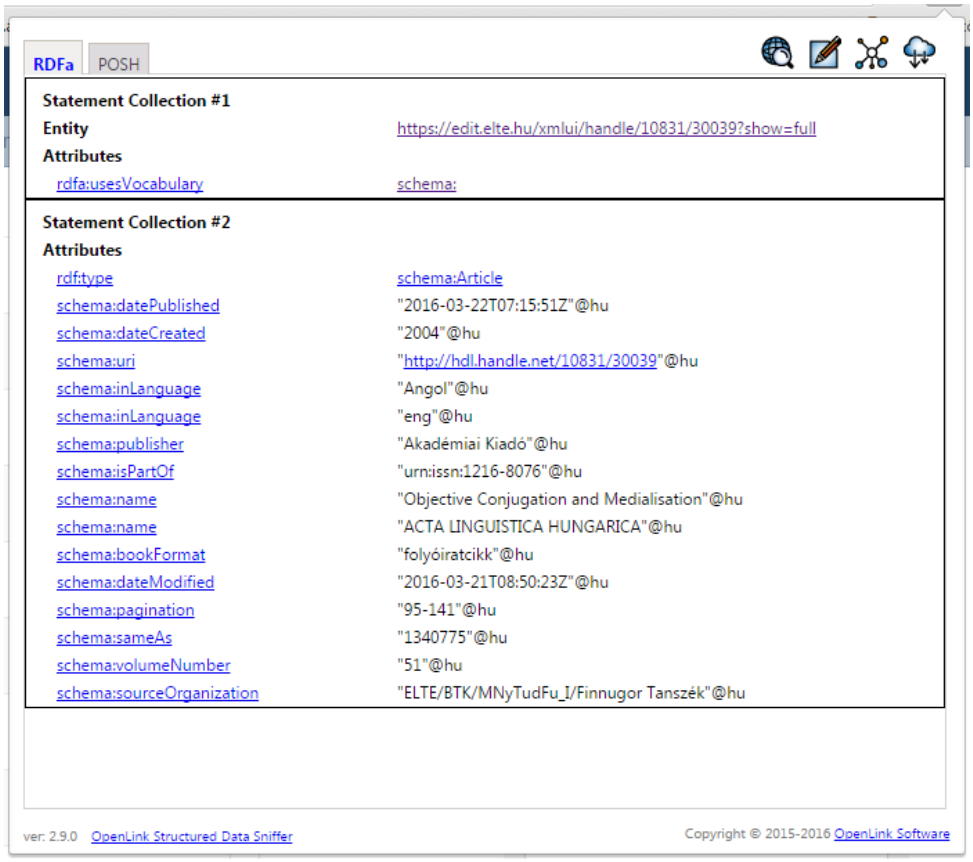


Fig. 4

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

Conclusion

In general, the major challenge of semantic applications in digital library environments is the way of representating the results to the end-users. The French semantic catalogue offers a really good example of getting new perspectives to information representation and cognitive reception. All the information about a

certain person, subject or work can be found with the help of corresponding links in a simple interface. Different editions and other cultural manifestations (theatre, music) of a work, author or person can thus be discovered. Moreover, the whole intellectual environment can gain the power of cultural inspiration through related people from theatre, music and other major cultural genres. One can start to browse a catalogue by looking for a certain piece of information and thus easily get lost just to discover the overwhelming richness of corresponding data. From a cognitive point of view, these new semantic catalogues offer a rather different perspective of information reception. Users meet with a complete information environment related to a certain subject with traditional and multimedia documents as well.

The German example is quite practical because it describes how a set of information about an emperor that could not be easily available on traditional web search engines or library catalogues can be stored and made retrievable. In a traditional library catalogue one can only search for a certain edition of a certain work of a certain author. In most cases the different editions of a certain work can be retrieved. However, all the other corresponding sets of data described above remain in a hiding position in the catalogue. The main power of the semantic cataloguing is that it connects different datasets that are described in standard format. In this sense, a new way of information retrieval is appearing that can be also relevant from cognitive perspective.

The Europeana example offers us a really good insight in the comprehensive efforts in order to aggregate resources from different semantic vocabularies, and highlights the essential need of collaboration in order to create high quality datasets that can be imported accurately by external partners.

In my opinion, BBC is the main actor of a silent revolution in linked open data field in the UK. They have built up a comprehensive linked data network and are currently running projects and building partnerships in order to share their datasets with cultural heritage institutions and other stakeholders. Moreover, they have initiated the establishing of an open Platform (RES) that makes media content and digital objects, documents of public collections available in a single environment. We can realize that the border between public collections, data and content providers is really flexible in this sense. The integration of data and content for the benefit of the society really matters. By creating new broad service environments it helps to discover different segments of cultural heritage.

By using semantic mark-up tags based on the HTML5 standard it has become possible to put semantic elements to virtually any homepage. Retrieving web-based information can be more comprehensive and complex in this way.

Search attitudes of people can be rather different in a linked data environment when compared to related traditional resources. The design and functional representation of semantic catalogue and application interfaces can be rather important: users have to be supported to make proper search terms. Representing and curating corresponding datasets in such a way that one would not get lost among different information resources and could easily decide how to go on in the searching process is a real challenge. Current web search engines (like Google) started to use the information datasets complemented with semantic markup tags in an intensive way.

The last major point of searching is focused on the quality assurance of new type of semantic catalogues and content applications. New indicators, benchmarking elements have to be implemented in order to enhance the quality criteria of these systems. The user-centred benchmarking focus, the use of cognitive analytical components related to user experience factors, seem to be really relevant in this context.

List of References

BBC, 2016a. *BBC ontologies* [online]. [cit. 2016-01-06]. Available from: http://www.bbc.co.uk/ontologies

BBC., 2016b. *BBC RES project* [online]. [cit. 2016-01-06]. Available from: https://bbcarchdev.github.io/res/

BBC, 2016c. *Linked data projects-BBC News LAB* [online]. [cit. 2016-01-06]. Available from: http://bbcnewslabs.co.uk/categories/linked-data/

BBC News Lab, 2016a. *BBC Juicer API* [online]. [cit. 2016-01-06]. Available from: http://bbcnewslabs.co.uk/projects/juicer/

BBC News Lab, 2016b. *WAT application-BBC News Lab*. [cit. 2016-01-06].

Dataset from Deutsche Nationalbibliothek (Joseph II.), 2016 [online]. [cit. 2016-12-18]. Available from: http://d-nb.info/gnd/118558404

BISHOP, H., 2016. *RES blog-Review of the year* [online]. [cit. 2016-01-06]. Available from: http://res-project.tumblr.com/post/154758429738/a-review-of-the-year

Getting started with schema.org using Microdata, 2016 [online]. [cit. 2016-01-01]. Available from:http://schema.org/docs/gs.html

GODBY, J., 2017. *Library Linked Data in the cloud: from proof of Concept to Action* [online]. [cit. 2017-01-15]. Available from: https://www.asist.org/Webinars/Webinar-01-12-2017-12344.pdf

HORVÁTH, Á., 2010. *Linked Data at the National Széchényi Library: road to the publication* [online]. [cit. 2016-12-18]. Available from: http://swib.org/swib10/vortraege/swib10_horvath.ppt

HORVÁTH, Á., 2016. *RDFa - schema.org: unity of document and semantic web*. Available from: https://conference.niif.hu/event/5/session/10/contribution/27/material/slides/0.ppt

MANGUINHAS, H., CHARLES, V., ISAAC, A. and T. HILL, 2016. *Entitifying Europeana: Building an ecosystem of networked references for Cultural Objects* [online]. [cit. 2016-01-01]. Available from: http://www.slideshare.net/HugoManguinhas1/entitifying-europeana-building-an-ecosystem-of-networked-references-for-cultural-objects

MCGINNIS, J., 2016. *The Web as a CMS: How BBC joined Linked Open Data*. Available from: http://ontotext.com/the-web-as-a-cms-how-bbc-joined-linked-open-data/

MEEHAN, T., 2014. The impact of Bibframe. In: *Catalogue & Index*, (177), 2–16. Available from: http://search.ebscohost.com/login.aspx?direct=true&db=lih&AN=110055753&site=ehost-live

National Library of France, 2014. *About data.bnf.fr* [online]. [cit. 2016-01-06]. Available from: http://data.bnf.fr/about

POWELL, J. E. et al., 2011. Graphs in Libraries: A Primer. In*: Information Technology and Libraries* [online]. **30**(4), 157–169 [cit. 2016-12-18]. Available from: http://search.proquest.com/docview/905719539

Resource Description Framework (RDF), 2014. *W3C Semantic Web* [online]. [cit. 2015-12-16]. Available from: http://www.w3.org/RDF/

Semantics conference Linked Data Award- BBC RES, 2016. Available from https://2016.semantics.cc/projects/research-and-education-space-res

Victor Hugo Les Miserables- semantic dataset, 2016. Data. Bibliothéque nationale de France [online]. [cit. 2016-12-18]. Available from: http://data.bnf.fr/13516296/victor_hugo_les_miserables/

Summary

WIDENING THE LIMITS OF COGNITIVE RECEPTION WITH ONLINE GRAPH DATABASES ON THE SEMANTIC WEB

Márton Németh

The main aim of this paper is to provide examples of the possibilities to extend the limits of cognitive reception via online graph databases related to different semantic web projects in public collection environment. Following the short general introduction of the relations of semantic web and public cultural collections, some current implementation projects and tools are being reviewed with the focus on public collections (library or library-related) interfaces using semantic web tools. Furthermore, the paper focuses on some benefits and challenges of the semantic web based solutions and the reception of these new kinds of interfaces by online users. However, semantic web based on machine-to-machine communication offering appliations towards human end-users is an essential task. Libraries, in cooperation with other stakeholder partners, can play a primary role in this field.