From Inventory Book to Expert System

Data Cleaning, Data Enrichment and Semantic Technologies in the Archaeological Collection of the Wosinsky Mór County Museum – A Case Study

ANDRÁS SIMON, Qulto – Monguz Ltd.
ENDRE FÜLÖP, Qulto – Monguz Ltd.
MÁRTA VÍZI, Wosinsky Mór County Library, Szekszárd

Though the amount of digitalized information is constantly growing, most of it is hidden in the “deep web”, which means that users are unable to find them. A professional solution for this problem is a cloud based Integrated System, a semantic network of identified, qualified and tagged metadata and full text or visual information, originating from databases and repositories of archaeology, digital inventories and catalogues, making browsing, search and displaying the linked data on the Web possible. Data is coming from databases of several types created for various purposes. It is prepared by local experts as data masters, and is linked by the managers of the collection management and knowledge organisation systems. Such data cannot be regarded as database records anymore, but rather statements acting as nodes and links of a semantic network. Linking all kinds of data nodes with each other, the system becomes the network of semantic statements, classifying the nodes and the link joining them as well. The innovation of this system is the integration of statements of various data masters, making the reuse of recorded metadata possible. The system prepares the search indexes, synchronises the elements of various data sources, cares for the protection of non-public data, and displays the relevant information on the web interface. This article may be helpful for heritage institutions that are about to implement an information system.

Key words:
Expert system, Semantic web, Digitizing.

CHNT Reference:
András Simon et al. 2019. From Inventory Book to Expert System - Conference Poster.

INTRODUCTION

By using the Qulto product portfolio, integration of archaeological inventory data, museum inventory books, online and collection catalogues become possible. The biggest advantage of such integration is that digitization of data has to be done only once, and then, in parallel with keeping master data, data cleaning and enrichment can also happen. The main goal of the integration and data enrichment, however, is to reconstruct the expertise of qualified specialist of the domain and hence to build a system that provides knowledge on a level similar to that of human experts, that means to create a domain expert system. Qulto has several museum clients in Hungary, including the Wosinsky Mór County Museum in Szekszárd (Tolna County Hungary). Their archaeological and inventory data taken from the museum catalogue undergo thorough selection and data enrichment after which they are automatically sent to the shared aggregated database of Hungarian museums, to MuseuMap, and finally to Europeana.

Aims, methods, tools

There are millions of data elements in digitized inventory books and other documents of museums, but as being part of the so called “deep web”, they are almost completely hidden from internet users. The information included in these digital inventory books and catalogues of Collection Management Systems are almost invisible and

Authors addresses: András Simon, Customer's manager: Monguz Ltd. 6072 Szeged Jobb Fasor, 6-10 Hungary; email: asimon@monguz.hu;
Endre Fülöp, It business analyst: Monguz Ltd. 6072 Szeged Jobb Fasor, 6-10 Hungary; email: efulop@monguz.hu

inaccessible for the tools of external discovery systems and search engines. Our R&D project aimed to find the possibilities of preparing the public data elements of the digitized inventory books of museums, to be the part of the "surface web", as a valuable information resource for the researchers and for the public, as well.

Preparing catalogue items as useful information for the surface web means not only that we migrate the inventory data to an open access catalogue, but also that we have to make the elements of the metadata searchable and understandable for the public, and for the experts as well. The expressions used in the database have to be identified and qualified both in a multilingual and an interdisciplinary environment. When dealing with the descriptions of archaeological objects and inventory items of museums in the archaeological collections, the content managers of databases have to work the most with the nominations, the geographical terms and the periods.

This case study is going to report about this R&D project by presenting some records from the digital catalogue of the archaeological collection of Wosinsky Mór County Museum, to show problems and their solutions (Fig. 1). For data cleaning, data enrichment, and publishing the results on the semantic web, the applications of Qulto Ltd. for museums were used. Some of the illustrations are taken from the online web catalogue of MuseuMap web interface which is the aggregating portal for Hungarian museums, and part of our references as being developed by Qulto [Vízi 2015].

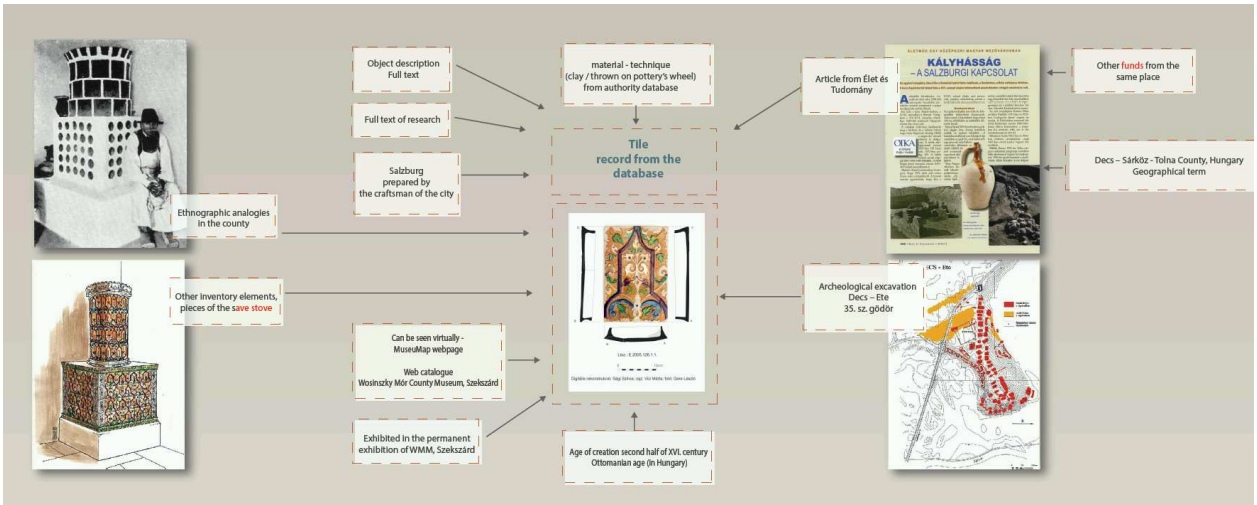

Fig. 1. Semantic network of different entities: Archaeological location, picture of a reconstructed object, archaeological fund as inventory item of a museum, bibliographic record of an article, other attached information.

De-duplicating of data

The digitized records of inventory books generally have a plenty of variations of the same words due to mistypings, errors, lack of conventions and standards of the work, and the changes of regulations and customs in museums. Such duplicates can occur even if the Collection Management System of the institution uses dictionaries or authority lists. If one makes a search, he has to enter all variants of a word to get all the relevant results; otherwise the search will not be comprehensive. Therefore, the first step of data cleaning is running a semi-automatic process of the Qulto ICMS in order to put the various strings for the same words into canonical form.

For example:
For a human being the expressions:

6th century
6th c.
VI. c.
VI. century
6th CE
Ad VI. century

have the same meaning, but for a search engine these strings are different. An expert, as the manager of the system, has to prepare (or choose) an eminent one, for example:6th century (AD 501-600)

Synonyms, variants of names

The next step of preparing an expert system is to identify the synonyms. It requires domain expertise to do so, as the focus is shifted on the semantic connection between the various expressions (semantic), instead of the way of writing (syntactic). For the same period, location, or technique more than one expression can be used. An expert system is required to know all expressions used for the same concept, and also to have information about the semantic relationship between them. Thus, in the second phase of the project we prepared a thesaurus (unified name-variant) database from concepts of the domain model of the archaeological collection (e.g. periods, materials, locations etc.). Each record of this database contains every known synonym of a given semantic content.

For example:

Lombard era (or Longobard era as a synonym)
AD 518-568
Late Gepidic era
Mid VI. century

The expert has to choose an eminent one here too, e.g.: Lombard era (AD 518-568)

For an expert of history or archaeology, the strings are the same, but only in the Carpathian Basin. (The Lombard era begins in Italy in 568). In addition, names of countries and folks can also have different words, expressions, and also the same word can have several meanings, especially in different languages. For example, in Hungarian, Lombard and Longobard are not synonyms. Lombard means the habitant of the Italian territory Lombardy, and Longobard (or Langobard) means the Germanic folk of the early Middle Ages. In Hungarian,there are two different words for the concept “Italy”, one for the historical area and one for the modern Italian state found in AD 1860.

Building a multi-hierarchical namespace

A multi-hierarchical namespace serves to represent the subordinate relationship of the terms and concepts of some special field. A period can contain another shorter period, accordingly a location or an ethnic group can involve a smaller area or community. There should be information about these relationships in an expert system in order to let the search processes be enhanced according to the relationship between a concept and the one which is part of it. In the third step we have built a multi-hierarchical namespace by using the thesaurus of the authority records that have been created from the data in the digitized inventory books of the archaeological collection. All the items are relevant which are linked to a concept of the database, which are subordinated to a selected expression of the namespace.

If we want to link the period Lombard era (in the Carpathian Basin) to a broader term, we can use the Migration period, which begins in the Carpathian Basin in AD 378, (The battle of Hadrianopolis) and ends in 805 AD (the Frank /Carolingian conquest).

This broader term can have several narrower terms:

Lombard era (AD 518-568)
1st half of bth c. (AD 801-850)
AD 788
Hun era (AD 420-455)
early Avar era
6th century (AD 501-600) etc.

By linking the records to each other, we can build a thesaurus to prepare a graph for the periodical terms [Bánki et al. 2016].

Entering the semantic web, using ontology

Only a part of the expressions' relationships of a discipline can exist in a single namespace database, which is not suitable for manifesting professional knowledge to its fullest. However, with the spread of semantic technologies it becomes possible that the units of various namespace management systems relate to each other to share their knowledge. This helps us to extend the relevant knowledge of our systems on such a level that we could not imagine before. As the 4th step of the project, we linked the records of the catalogue to a new generation semantic network of the internet so that we become able to use the databases of Geotaurus and Wikidata, to provide relevant information to researchers. Using semantic technologies, offers novel possibilities even within the system itself: by building domain, we can prepare algorithms for associating between expressions, and for automatic data enrichment. By using units of other knowledge management systems, and the automatic associations prepared by a built-in ontology (i.e. rules for applying domain knowledge to inferencing and problem solving), the search engine of the system can give new relevant results for the customer.

The periodical term Lombard era (AD 518-568) can have related terms:

Race – Germanic
Nation – Lombard /Longobard
Geographical term – Hungary

The Lombard nation can be linked to a race, Germanic, and to this race other folks, for example Gepid, or Alemann can be linked. Also, the geographical term Hungary can be linked to the Roman Province Pannonia, and to the Carpathian Basin. Having the network of the concepts, the customers can use the linked elements as related search terms, therefore more relevant results can be reached.

These related terms can exist in the database of the collection management system of the institution, but they can also come from external data sources, for example from a database of the names of animal or plant species [Simon 2017].

From the Catalogue to the Semantic web

A sophisticated database of archaeological objects and items can integrate the digitized records of museum inventory books and library catalogues, too. In this case, it also has to contain items of bibliographies, authority records of nominations, geographical terms, personal and corporate names. If database also includes dictionaries of periods, animals and plant species or subject headings, it can only exist in a semantic network with identified, qualified, and mutually linked element. For this purpose, special applications have to be developed where these data can be stored and managed. The records themselves can either be imported from the collection management systems or entered directly in an expert system. From the database of an expert system, the metadata can be published to the semantic web [Ungváry 2012].

Conclusion

By preparing a semantic network from various databases, we perform the above steps in the case of each and every customer who purchases a CMS from the Qulto portfolio. By manipulating millions of data elements, local semantic relationships are built in a semi-automatic way for all kinds of collections. Therefore, museum catalogues are converted into linked networks of records. These network nodes can consist of inventory items of museums, items of library catalogues and bibliographies, authority records for personal names, corporate, geographical locations, names of species, and expressions for the museum catalogues, for materials, techniques, subject headings, and various classification marks, and also media records of full text data, images, sound materials etc. Such a system can help both the experts of the institutions and the customers not only to use it as a vocabulary, but in the future an application of artificial intelligence can be based on it, as well.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the Wosinsky Mór County Museum's kind contribution to the poster presentation and the article based on that.

REFERENCES

Zsolt Bánki, Tibor Mészáros, Márton Németh, András Simon. 2016. Checking the identity of entities by machine algorithms is the next step to the Hungarian National Namespace. In *Code4lib Journal 2016*.

Márta Vízi.2015. Kályhásság – a salzburgi kapcsolat. Életmód egy középkori magyar mezővárosban. In *Élet és Tudomány* Vol.52. (2015) 11. sz.

András Simon. 2017. Rekord vagy háló – tudásreprezentációs eszközök személynévtér-építéshez gépi katalógusokból kinyert adatok alapján Névtér: új fogalom régi megoldás. In *Tudományos Műszaki Tájékoztatás* 2017.10.

Rudolf Ungváry. 2012. A névterek és az adatok tulajdonságai. In *Tudományos Műszaki Tájékoztatás* Vol.59. sz.

András Simon, Endre Fülöp, Márta Vízi.2017. From inventory book to Expert System. Conference Poster CHNT 2017 November 8-10 Vienna.

Imprint:
Proceedings of the 22nd International Conference on Cultural Heritage and New Technologies 2017.
CHNT 22, 2017 (Vienna 2019). http://www.chnt.at/proceedings-chnt-22/
ISBN 978-3-200-06160-6

Editor/Publisher: Museen der Stadt Wien – Stadtarchäologie
Editorial Team: Wolfgang Börner, Susanne Uhlirz
The editor's office is not responsible for the linguistic correctness of the manuscripts.
Authors are responsible for the contents and copyrights of the illustrations/photographs.