IDENTIFYING SEMANTIC RELATIONSHIPS IN DIGITAL LIBRARIES

Zoltán HARSÁNYI, Viera ROZINAJOVÁ

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
{harsanyi,rozinajova}@fiit.stuba.sk

Abstract. In this paper we deal with the interoperability of digital libraries concerned with identification of hidden relationships within various data sources. By means of semantic processing and reasoning techniques we attempt to find answers to sophisticated questions which are sometimes difficult also for human experts. Our initial interest was to analyze data from museums, where we have found interesting information concerning artists, their life and work. By analyzing enriched data, we could identify semantic relationships among the records, which can help us understand how these artists were influenced by each other. This paper summarizes our results achieved in previous research.

1. Introduction

One of the great challenges in digital libraries concerning cultural heritage is the problem of semantic interoperability. The information about artifacts would be much more valuable if the machines could "understand" the meaning of digitized information and various sources of information could be interconnected in order to acquire complex knowledge about certain domain. In this paper we propose one approach to solving the problem of semantic interoperability in the domain of cultural heritage. Our interest was to analyze records from art museums and national libraries of several countries, which contain considerable information in the form of metadata about various persons as artists [1].

1.1 Project VIAF

In order to address the issues mentioned above, project called VIAF was created. Project VIAF (The Virtual International Authority File), implemented and hosted by OCLC (Online Computer Library Center), is a joint project of several national libraries and selected regional and trans-national library agencies (22 agencies in 19 countries).

Using VIAF API we can search for authority data by keyword, local name, preferred name form, title, source, or control number of source records and retrieve authority records in different formats from several national libraries.

1.2 Project WorldCat

OCLC and its member libraries cooperatively produce and maintain WorldCat, the world's largest online database for discovery of library resources. Member librarian institutions of WorldCat are dedicated to providing access to their resources on the Web, where most people start their search for information.

WorldCat also provides different Web services, called OCLC Web Services, which represent a set of tools and APIs that expose provided data to OCLC members and partners.

2. Related Work

Cultural heritage and digital libraries are complex, closely related disciplines that usually benefit by applying and using new, innovative technologies. Semantic interoperability is a top-priority issue for cultural heritage applications, due to the need to aggregate and disperse data within knowledge aware services.

The main challenges for the semantic digital libraries are:

- Integrating information based on different metadata
- Providing interoperability with other systems (not only digital libraries) on either metadata or communication level
- Delivering more robust, user friendly and adaptable search and browsing interfaces empowered by semantics

Metadata play very important role in digital libraries, as they enable higher quality of "understanding" stored data. The architecture of the semantic digital library takes into account not only the legacy metadata descriptions of the resources but also annotations provided by the community of users and the information about the users themselves. Services offered by semantic digital libraries should now assist users in efficient discovery techniques in the new, interconnected information space.

3. Data processing and the method description

The main part of research is to design and implement a method for obtaining data, enriching them with additional information and analyzing the semantic relationships among them using formally defined rule based mechanism. The procedure, which we designed, could be summarized in following steps:

- Obtaining the data from one of the national libraries and processing them
- Requesting for information from VIAF Web service for each record
- Requesting for information from Web services of WorldCat
- Normalization and data de-duplication
- Analyzing the relationships among the records and creating reliance graphs

– Visualizing the records and providing a user interface to allow users to search for information in the records and also get a possibility of browsing graphs

We have gathered 800 records about different artists in XML (MARC21) format. After the processing of these data we used other Web services to enrich our basic records (WorldCat Search API, xISBN and WorldCat Registry).

We analyzed and identified different relationships, which we categorized into the following main groups:

– Identified semantics: family relationships, school – person (teacher, classmate), employer – person (superior, colleague), etc.
– Obtained relationships: works concerning the artist, relationships to organizations, aggregated information (e.g. the number of painters), etc.

The final stage of the project was about the effective visualization of different records and the identified relationships. We designed different user interfaces, one for displaying enriched records, another for visualizing aggregated relationships using charts or tables and for visualizing dependencies using graphs.

3.1 Identifying cooperating associations

Our records contain information about memberships of artists in different organizations and associations dealing with the common issues of art. We wanted to find out which of these organizations were/are likely to cooperate. So we first filtered the ones which had at least 3 members. The most "popular" organizations from our records have 67 members. After performing this process we created a network of pairs of organizations which have common members, i.e. artist a_1 is a member of organizations o_1, o_2 and artist a_2 is also a member of these two organizations. We filtered out those pairs of associations which have less than 4 common members. We found 2 organizations having 17 common members. After this process we created a network of these associations (Figure 1) (only initials of these organizations are shown in the graph), visualizing the possible co-operations among organizations. Let us describe the process by example. The first organization in Figure 1 is SVUM, the first one on the second level on the left side is UBS and the one below SVUM is SVUVB. SVUM and UBS have 17 common members among all artists in our dataset. SVUM and SVUVB have 7 common members.

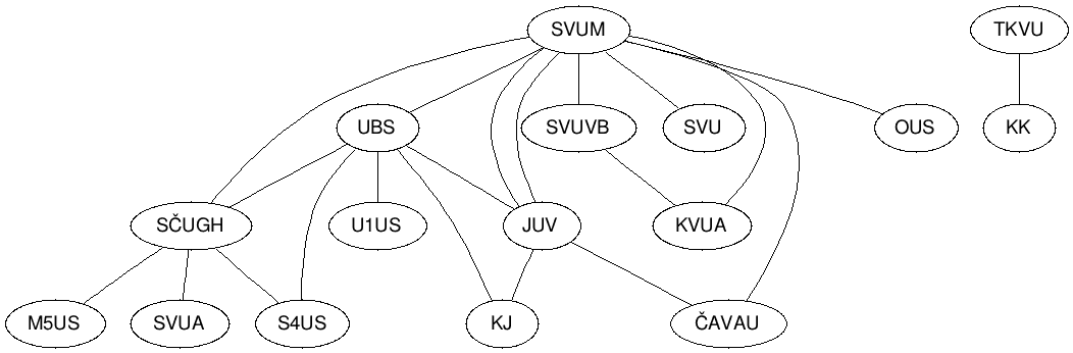

Figure 1. Cooperation graph of organizations having 4 common members between 2 of them.

We analyzed this network deeper and checked those triples of organizations which have at least 3 common members, i.e. artist a_1 is a member of organizations o_1, o_2 and o_3, and artist a_2 is also member of these three organizations.

We would like to know which artists influenced the organization in significant way. So we defined a rule to find these artists. An artist or an organization can be in two different relationships with an art-work (book, publication, etc.). The first one is the authorship (artist is the author of the work) and the second one is a case when the work is about the artist or organization. So if we analyze common members of an organization and find any works of these members dealing with the organization or other colleagues, we can assume that these members or institutions are/were significant. Another case is when artists, belonging to common organizations, deal with these organizations in their work. Then we conclude that the membership in this organization was quite significant for them and could influence their work considerably. After applying these rules, we identified four organizations where a "significant member" has been found. These organizations are SVUM, UBS, SČUGH and SVUA.

4. Conclusions

In this paper we have presented a way of enriching and identifying relationships among bibliographic records (in the domain of digital cultural heritage). Our approach is based on data from art museums and national libraries, enriching them with different data and analyzing the possible connections among the given records. We have also developed a tool for visualizing these records and relationships between them, which is intended to allow users to explore the records.

We ascertained that creating and applying different rules on the set of enriched data can lead to identification of hidden relationships between them. In this paper we described one of the possible cases of their evaluations. This can be quite significant result from the point of view of moving towards more "intelligent" mechanisms of retrieving and processing data in digital libraries not only in the domain of cultural heritage.

Acknowledgement: This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0208-10, the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11 and by Research & Development Operational Programme ITMS 26240220039, co-funded by the ERDF. The authors also wish to acknowledge the generous support of OCLC for providing no-charge access to the WorldCat database in support of this research project.

References

to other papers publishing the results that are summarized here

[1] Harsányi, Z., Rozinajová, V., Andrejčíková, N.: Identifying semantic relationships in digital libraries of cultural heritage. In: EuroMed 2012: Progress in Cultural Heritage Preservation, LNCS 7616. Springer, Berlin / Heidelberg, (2012), Ch. 78, pp. 738-745.