A framework for interacting FRBRoo/ CIDOC CRM data: With emphasis on heterogeneous needs

Maliheh Farrokhnia
Dept. of Archivistics, Library & Information Science
Oslo and Akershus University College of Applied Sciences, Norway
Maliheh.Farrokhnia@hioa.no

ABSTRACT

Considerable time and efforts have been spent on developing standards and models that provide interoperability between information system in archives, libraries and museums. Among them CIDOC Conceptual Reference Model (CRM), primarily targeting museum information and later FRBR object-oriented model (FRBRoo), which is an extension to CIDOC CRM developed to integrate information across memory institutions (such as archives, libraries and museums) in a semantic way. Despite different efforts to implement these conceptual models in integrated systems with a substantial amount of cultural heritage information, information retrieval and its representation still has some challenges. The author suggests that information visualization can be applied to display better the semantic structures and relationships between objects described by FRBRoo/ CIDOC CRM. This research also presents the potential of graphical information visualization as a knowledge network and proposes a framework about how a user interface in such integrated systems could be semantically established according to the user needs. Application of this framework can easily be used in many integrated information systems.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *graphical user interfaces, user centered design*;

General Terms

Design, standardization

Keywords

Information Integration, Cultural Heritage, Information Visualization, Conceptual Model, FRBRoo, CIDOC CRM, User-Centered Design Patterns, Graphical Visualization, User Studies

1. INTRODUCTION

During the last few years there has been an increased focus on interoperability and data integration across archives, libraries and museums. As systems have become more complex, retrieving and accessing relevant information efficiently has become more difficult, partly because of the nature of modern information systems, which tend to rely on vast quantities of stored data with a high degree of structural complexity [13]. According to Vassallo [18], topic maps or at least the concept of a net of relationships, independent from the level of the occurrences, allow the description of a single object to be carried out in conformity with the specific descriptive standard, but at the same time they create a net that enables the integration of various objects.

With the increasing efforts to provide integrated access to memory institutions (such as archives, libraries and museums) considering differences in organizing information and the need to integrate data sets, object-oriented reference models such as CIDOC Conceptual Reference Model (CRM) and FRBRoo as an extension to CIDOC Conceptual Reference Model (CRM) have been developed. The models aim to integrate information across these memory institutions in a more effective way than with current approaches.

These reference models can be considered to represent a significant shift in the semantic metadata interoperability achievement of memory institutions and cultural heritage. Besides, they provide a detailed model of how all these materials about one or more performances, found distributed over archives, museums and libraries are related based on the underlying common history [3].

Although such rich conceptual models can improve the semantic relationships between information objects, their implementation is still a challenge. In addition to being complicated to implement, the representation can also be complicated to understand for end-users.

For many years, much effort has been made to overcome the shortcomings of online catalogues related to the uncontrolled retrieval and display. For instance, Fattahi [4] proposed the concept of super records in order to make the retrieval more meaningful and manageable so that the relationships of dependence and subordination, of similarity and difference between or among related entities would be more clearly demonstrated. Moreover, hypertext links which form the main structural paradigm of the Web at present, allow one to navigate manually in a network of information. But these links will never form a global semantic network [2].

It seems that the graph data structure is becoming a more accepted representational medium and as such, may soon displace the linked table data structure of the relational database model [12]. Knowledge Graph is a functional enhancement that attempts to provide actual information about the subject of the user's query rather than just a list of links. It aims at not just providing the answer to what you asked, but also the answers to the questions you probably should have asked [5]. Knowledge Graphs are linking information together in a meaningful way and presenting the integrated results to the user. In cases in which corresponding data elements have inherent relationships, graph visualization methods are commonly applied to support the better understanding [15]. The data can be represented by the nodes of a graph, with the edges representing the relations. Although graph visualization techniques are widely used in many application domains, they have some limitations that have to be dealt with. The size of the graph is a major issue in systems with huge amount of information. Large graphs can pose several

Digital Libraries 2014 Doctoral Consortium
September 8, 2014, London, UK.

difficult problems. Even if it is possible to layout and display all the elements, the issue of viewability or usability arises, because it will become impossible to discern and navigate between nodes and edges. Consequently, a first step in the visualization process is to reduce the size of the graph. After discovering clusters in the data, we can reduce the number of elements to display by restricting our view to the clusters themselves [7].

Then, to have a better integrated system, it is necessary to have a user-centered display. Understanding the user's intent or information need that underlies a query has long been recognized as a crucial part of effective information retrieval. The point is that the catalogues may contain more than the user may be expecting. Users' needs vary a great deal. While some users may find any edition of a work useful, others may require a specific edition with a particular feature. There are also users who look in the catalogue for a particular manifestation of a work or a work in a particular format [4]. This variety will be more extended in the integrated cultural information systems which are supposed to offer information services semantically to the heterogeneous users from archives, libraries and museums. So, in order to develop a system that will manage the relationships between different areas of cultural heritage, it is necessary to manage entities of various nature (for example, classes of objects as fonds, works, their creators, publishers, and rights owners.) [18]. Then we can offer a proper clustered information representation considering users' needs and goals in different domains. In fact, without a semantic representation, even a well-designed information architecture is invisible to users.

This research will propose a framework for interacting with FRBRoo/ CIDOC CRM data, with its increased emphasis on heterogeneous cultural information needs of users of archives, libraries and museums. The framework have the potential to enhance the semantic information seeking and retrieval in an integrated system. The result of this research will lead to a useful interaction between the archives, museums and libraries as data provider institutions and the users with different needs and interests as data consumer or even data provider.

2. LITERATURE REVIEW

Semantic data integration has been a dynamic and challenging research area for many years, pursuing the goal to provide users with a uniform interface to access to the semantic related data. In 1997 Semmel and his colleagues developed a framework based on intelligent agents and distributed objects that will facilitate the fusing of new legacy systems, interoperability, and the creation of intelligent interfaces. In conceptual modeling schemes based on basic entity-relationship (ER) constructs, data are represented by classes and relationships among those classes. Here, there is insufficient knowledge to infer semantically reasonable queries over complex domains [10]. To counteract problems with traditional Universal Relation (UR) approaches, the Applied Physics Laboratory (APL) has developed a system known as QUICK (QUICK is a Universal Interface with Conceptual Knowledge). This system uses extended entity-relationship (ER) design knowledge to facilitate query formulation and meaningful information retrieval across semantically related heterogeneous databases [13, 14].

Doerr and Iorizzo in 2008 offered a new approach to deal with the problems of semantic interoperability and the creation of a global knowledge network by presenting a powerful information architecture that could emerge based on the CRM, as a generic global ontological model based on relations and co-reference rather than objects, for summarizing, structuring, and combining existing data. They also offer semi-automatic maintenance of co-reference links, and public engagement in the creation and development of the network [2].

To support the user in exploratory search over a large number of items Urruty and his colleagues [17] introduced an interface for video retrieval that automatically presents suggestions by extracting textual and visual features of relevant shots. They proposed clustering of the retrieved results based on low-level features to create groups of similar content. However, this fixed set of features would not be useful for heterogeneous multimedia objects.

In a technical report published in 2011, Doerr and Tzomanaki proposed a new framework for querying semantic networks: For formulating queries, the user is presented a small list of configurable "Fundamental Relationships" (FRs) and relevant specializations, easy to comprehend, that abstract by rich deductions from an underlying semantic network of much more specialized metadata comprising explicit event descriptions. These FRs simulate a much simpler semantic network which covers as many generic questions as possible with a high recall. The specializations of the FRs enable systematic increase in the precision of queries on demand, down to the level of detail of the underlying network [16].

Among cultural heritage projects for integrating heterogeneous metadata CultureSampo has been developed since 2004 as a part of the FinnONTO project and its first public prototype was published in 2008. It is a system that addresses the semantic web challenge of aggregating highly heterogeneous, cross-domain cultural heritage collections into a semantic intelligent system for human and machine users [9].

Another related project is Europeana which launched in 2008 with the goal of making Europe's cultural and scientific heritage accessible to the public. The Europeana Data Model (EDM) attempts to structure and represent data delivered to Europeana by the various contributing cultural heritage institutions. This model facilitates Europeana's transition from a closed data repository to an open information space that integrates with the Web architecture and the Linked Data principles for identifying and exposing resources on the Web [6].

Recently, as part of the SYNAT[1] project, different metadata formats used by Polish digital cultural heritage institutions were mapped to CIDOC CRM and FRBRoo to provide interoperability. As a work of mapping MARC 21 to FRBRoo, Mazurek and his colleagues emphasized on the role of FRBR conceptualization in order to facilitate easier understanding of the structure of library resources. Besides, they confirmed the FRBRoo ontology as an interesting solution to apply in repositories where museum and library resources metadata are combined. However, they discuss some challenges in automatic translation of existing digital library metadata records such as the lack of a clear line between the four FRBR levels or between physical and digital resources. Another challenge they mentioned was how to present such structured information to digital repositories end-users [11].

In another project in 2012, Chen and Ke [1] used FRBRoo as a shared ontology to integrate the heterogeneous metadata generated by museums and libraries. Based on their findings, they emphasized the inter lingua role of FRBRoo in aligning museum and library metadata to achieve heterogeneous metadata

[1] SYNAT is a research project aimed at the creation of universal open repository platform of networked resources of knowledge for science, education, and open society which is funded by the National Center for Research and Development (grant no SP/I/1/77065/10).

integration and semantic queries without changing either of the original approaches to fit the other.

In recent years, different works have been done in the area of information integration of cultural heritage information. In order to complete them and make them more useful for end users, a semantic interface based on FRBRoo/ CIDOC CRM models according to users' preferences will be proposed in this research.

3. RESEARCH QUESTIONS

In conducting this research the following questions are addressed:

1. What are the users' predefined understanding (conceptualization) of resources held in archives, museums and libraries?
2. How do users' conceptualization in different domains (libraries, museums, archives) match to the FRBRoo/ CIDOC CRM models?
3. How do users expect to search and explore FRBRoo/ CIDOC CRM data in libraries, museums and archives?
4. How should a user-centered framework for interacting with FRBRoo/ CIDOC CRM be designed?

4. OBJECTIVE

This work is designed to find an alternative framework for interacting with FRBRoo/ CIDOC CRM data considering users' needs and requirements in different domains of memory institutions.

5. RESEARCH DESIGN AND METHODOLOGY

In order to determine how the information can be displayed per result in a semantic way in integrated databases based on FRBRoo/ CIDOC CRM, we need to get a better understanding of what users actually want from such databases.

A series of semi-structured in-depth interviews will be conducted with users in different domains of archive, library and museum in order to explore heterogeneous users' needs and preferences towards an integrated cultural information system.

For this stage, purposeful sampling is used in order to get most effective result. The potential participants in each of the three domains of memory institutions will be interviewed by the researcher. The number of interviewees may vary depending on when data saturation is achieved. Because of the probable participants with different nationalities, interviews will be conducted in English and each will last between 45- 90 minutes. The interviews will be voice recorded with the permission of interviewees along with note taking during the interview.

Through such in-depth interviews, the researcher is going to investigate the patterns in users' predefined conceptualization of resources held in memory institutions and also the heterogeneous users' preference and expectations form the result presentation in an integrated system.

The following questions will be asked to structure the interviews. The interviewees will be given some prepared queries and result lists including information resources with common topic or underlying background form different memory institutions:

- What is your understanding of different resources held in archives, museums and libraries?
- How you would organize the result list of this specific type of query? What are the factors that can influence your way of organizing the result list?
- Describe the entities or information elements that you would expect to find in a result list for this specific type of query?
- What kinds of entities or information elements do you prefer to see for each of these information resources in the result list?
- Which kinds of relationships do you think are important to be represented for each of these information resources?
- For what purpose or task are these kinds of relationships and entities useful and important?

As the result of this stage, the preferable core entities and relationships in the form of some semantic groups and patterns will be determined based on the users' conceptualization. In fact, the project will identify information needs, present and potential, similar and dissimilar related to each kind of information resource among the selected user groups of archives, museums and libraries.

For the next stage, the patterns of users' conceptualization will be matched to FRBRoo/ CIDOC CRM models in order to explore how these models can match the real users' conceptualization in different domains.

Finally, all the data collected for this research will be analyzed and the user–centered framework for interacting FRBRoo/ CIDOC CRM models will be designed and proposed.

6. DATA ANALYSIS

The process of analysis will be undertaken alongside data collection and a research journal will be kept in which emerging analytic ideas and hypotheses will be noted, developed and fed back into the data collection process.

7. ETHICS APPROVAL

Study procedures and research project will be approved by Oslo and Akershus University College of Applied Sciences and NSD as the Data Protection Official for Norwegian research institutions.

Although there won't be any risk to anyone in the research or anyone associated with it, all personal identifiers will be removed from the transcripts. Besides, the files of recorded interviews will be deleted 6 months after finishing the research.

8. SCHEDULE

a) Prerequisites (10 months)
   - Completing university prerequisites, PhD courses
   - Reading materials and developing theoretical framework

b) Data collection (8 months)
   - Developing data collection instruments and prerequisites
   - Conducting interviews

c) Data analysis (8 months)

d) Writing thesis (6 months)

e) Final revision (4 months)

9. REFERENCES

[1] Chen, Y.-N. and Ke, H.-R. 2013. FRBRoo-based approach to heterogeneous metadata integration. Journal of Documentation. 69, 5 (2013), 623–637.

[2] Doerr, M. and Iorizzo, D. 2008. The dream of a global knowledge network—A new approach. Journal on Computing and Cultural Heritage. 1, 1 (Jun. 2008), 1–23. . DOI = 10.1145/1367080.1367085.

[3] Doerr, M., LeBoeuf, P. and Bekiari, C. 2008. FRBRoo, a conceptual model for performing arts. Annual Conference of CIDOC, Athens (September 2008) (2008), 06–18.1838425657.

[4] Fattahi, R. 1996. Super records: an approach towards the description of works appearing in various manifestations. Library review. 45, 4 (1996), 19–29.

[5] Google's Knowledge Graph: Yeah, that's the Semantic Web (sort of): http://blogs.gartner.com/darin-stewart/2012/05/17/googles-knowledge-graph-yeah-thats-the-semantic-web-sort-of/. Accessed: 2014-08-20.

[6] Haslhofer, B. and Isaac, A. 2011. data. europeana. eu: The europeana linked open data pilot. International Conference on Dublin Core and Metadata Applications (The Hague, The Netherlands, 21-23 September, 2011), 94–104.

[7] Herman, I., Melançon, G. and Marshall, M.S. 2000. Graph visualization and navigation in information visualization: A survey. Visualization and Computer Graphics, IEEE Transactions on. 6, 1 (2000), 24–43.

[8] Hull, R. and King, R. 1987. Semantic database modeling: Survey, applications, and research issues. ACM Computing Surveys (CSUR). 19, 3 (1987), 201–260.

[9] Hyvönen, E., Mäkelä, E., Kauppinen, T., Alm, O., Kurki, J., Ruotsalo, T., Seppälä, K., Takala, J., Puputti, K. and Kuittinen, H. 2009. CultureSampo—Finnish culture on the Semantic Web 2.0. Thematic perspectives for the end-user. Proceedings, Museums and the Web (2009), 15–18.

[10] Markowitz, V.M. and Shoshani, A. 1989. Abbreviated query interpretation in extended entity-relationship oriented databases. Entity-Relationship Approach to Database Design and Querying (Amsterdam, 1989), 325–343.

[11] Mazurek, C., Sielski, K., Walkowska, J., Werla, M. and Supercomputing, P. 2012. From MARC21 and Dublin Core, through CIDOC CRM: First Tenuous Steps towards Representing Library Data in FRBRoo. CIDOC 2012. (2012).

[12] Rodriguez, M.A. 2009. A graph analysis of the Linked Data cloud. arXiv preprint arXiv:0903.0194. (2009).

[13] Semmel, R.D., Immer, E.A., Silberberg, D.P. and Winkler, R.P. 1997. Knowledge-based query formulation for integrated information systems. Johns Hopkins APL Technical Digest. 18, 2 (1997), 261.

[14] Semmel, R.D. and Silberberg, D.P. 1993. Extended entity-relationship model for automatic query generation. Telematics and Informatics. 10, 3 (1993), 301–317.

[15] Tarawaneh, R.M., Keller, P. and Ebert, A. 2012. A General Introduction To Graph Visualization Techniques. OASIcs-OpenAccess Series in Informatics (2012).

[16] Tzompanaki, K. and Doerr, M. 2012. A new framework for querying semantic networks. Proceedings of Museums and the Web 2012: the international conference for culture and heritage on-line (2012).

[17] Urruty, T., Hopfgartner, F., Hannah, D., Elliott, D. and Jose, J.M. 2009. Supporting aspect-based video browsing: analysis of a user study. Proceedings of the ACM International Conference on Image and Video Retrieval (2009), 47.

[18] Vassallo, S. 2006. Navigating through archives, libraries and museums: topic maps as a harmonizing instrument. Charting the Topic Maps research and applications landscape. Springer. 231–240.