Navigating through archives, libraries and museums: topic maps as a harmonizing instrument

Salvatore Vassallo

University of Pavia, Corso Strada Nuova 65,
27100 Pavia, Italy, salvatore_vassallo@tin.it

Abstract. The paper deals with the possibility of creating a topic map based system where different sectors of cultural heritage would interact with users, by monitoring the navigation histories of users and the statistics on the searches, in order to authorize variant form of names. The problem of managing different sectors and harmonizing them both from a structural and a semantic view point, by using topic maps, is also discussed. With regards to this, we are introducing two projects, which are largely based on the above mention use of topic maps.

1 Introduction

The paper considers use of topic maps in the area of cultural heritage from three view points:

- to manage the variant forms of a name, caused by the users' search itself. According to this, we carried out an analysis through questionnaires in order to test a hypothetical system built on this logics;
- to allow the management and the navigation through an archive: we will present a model finalized to the production of a guide for the exploitation of the archival fonds as well as the reorganization of the library, both owned by the "Archivio di Stato di Pavia";
- to navigate through archives, libraries and museums: using topic maps as a harmonizing instrument in conformity with the specific descriptive standards, but at the same time creating a logical framework enabling the interactions of various objects. This idea is at the basis of the CeDECA[1] project: a census about cultural heritage in the Oltrepò pavese.

[1] Centro di Documentazione Etnografica e di Cultura Appenninica, developed on behalf of Pavia University by Maria Antonietta Arrigoni, Federica Biava, Ester Bucchi de Giuli, Marina Chiogna, Paola Ciandrini, Elettra de Lorenzo, Elena Giavari, Flavia Giudice, Marco Savini and Salvatore Vassallo with the coordination of professors Pierangelo Lombardi and Paul Gabriele Weston.

2 Topic Maps and variant forms of names: a permanent renovation

In this case our study started from the analysis of the solutions adopted by products such as Aquabrowser[2]: the peculiar graphic layout of the latter showing the variant name options, led us to foresee the possibility of incorporating some of those functions into a topic map.

One of the aims of Aquabrowser (which, according to the scopes of our analysis is just an example) is to use the words related with the search to discover new information and to help users to formulate a new query. The discover function works like the associations in a topic map. The problem is that the software uses also the spelling variations (probably based on Levensthein distance ≤ 2) to determine the associations. Such an approach will necessarily cause a great deal of noise: for example, a search based on the string "Kenedi" (meaning Aaron Kenedi) will produce as an alternative form, the name "Kennedy". Another example is the case of "queen": here Aquabrowser uses as alternative form the term "queer" to generate other associations, which is quite obviously a problem.

Our idea is to overcome this limit, through a statistical analysis of the users' behaviors, in order to certificate the variant forms, no matter how they have been generated. For instance, if, among the average sample, a significant percentage of users research "Kenedy" and accept the option which is suggested, (i.e., the form "Kennedy") by selecting it and not leaving the page within the first 30 seconds[3], then "Kenedy" will be considered a variant forms of the name certified by the users, and will be included into the "Kennedy" topic (as variant or as basename) and used to generate the net of associations (as in Aquabrowser).

We have prepared a questionnaire with the aim of simulating users' approach to the research: five known personalities were indicated and the user was asked to write down how he would search each name into a hypothetical informative system. The test was carried out on famous people, but could have dealt with any other term (indeed, the idea of an automatic certification of variant forms of a name refers to any research term, even though it is undeniable that people's names seem to be among the most researched terms).

We tried to find an empirical formula to define the minimum rate to become a certified variant form: the main idea is to find an equation that decreases slowly when the number of questionnaires increases. In this meaning we analyzed in increasing groups the questionnaires, determining and testing, step by step, the minimum rate. The formula upon which the minimum rate varies according to the number of searches was calculated by interpolating such results.

---

[2] Aquabrowser, <http://www.medialab.nl/>, is developed by Medialab as a non conventional library OPAC interface. It appears like a system that allows the contextualization of terms, using a graphic environment comparable to graphic topic maps. Besides it offers the chance of navigating through variant name forms, trying to cater for accidental mistyping. It's indeed on this function that we based our first analysis.

[3] It is the time estimated so to exclude non profitable searches, evidenced by the quick leave of the page.

$$P = \frac{1}{k^{\log x}} \quad (1)$$

where P is the minimum rate, x is the number of questionnaires (in our case or the number of searches in the case of an information system) and k is a constant value (empirical range calculated between 2.0 and 3.0). This range is a consequence of the impact of the constant value on the inclination of the curve: in the presence of a highly homogeneous group of users one should decrease k to increase P and to refine the sample (for i.e. to exclude dialect form, typical of a homogeneous groups).

This solution and the equation now exposed were tested through a questionnaire filled in by nearly 600 persons, of different age and social extraction. So, with an average k = 2.5, according to the formula the minimum rate is 8%

Significant results that were obtained in relation to the above mention function were the following:

Table 1. Shakespeare – name form certified by users searches (≥8%)

Name form certified	Per cent of questionnaires
Shakespeare	82%
Shakespear	13%

Table 2. Krusciov – name form certified by users searches (≥8%)

Name form certified	Per cent of questionnaires
Krushov	50%
Kruscev	13%
Krusciov	13%
Crusciov	8%

Table 3. Beethoven – name form certified by users searches (≥8%)

Name form certified	Per cent of questionnaires
Beethoven	76%
Beethowen	11%

Table 4. Ceausescu – name form certified by users searches (≥8%)

Name form certified	Per cent of questionnaires
Ceausescu	32%
Ciausescu	29%
Chausescu	17%
Causescu	9%

Table 5. Tchaikovsky – name form certified by users searches (≥5%)[4]

Name form certified	Per cent of questionnaires
Tchaikovsky	6,5%
Chaicoski	5%
Tchaikowsky	5%

This idea could be integrated into a real system through the automatic analysis of statistic researches, thus certificating the variant forms of the name, according to users' "mistakes".

Undoubtedly a choice of this kind is laid open to criticism from the language purists' side, who could accuse our approach of laxity and of encouraging the language natural degeneration. Anyway our first aim is users' satisfaction and, in this case, the research success. If you better consider it, topic maps can turn into a didactic instrument, since – navigating through the variant forms of names (or, to better say it, through usual errors) – you can recognize and consequently avoid the most common spelling mistakes.

A system such as Aquabrowser, for example, can evolve, by showing through graphs only the options and the associations included in the topic maps (we could say certified by the users).

3 Navigating through an archive

In this paragraph our intent is to illustrate the possibility of creating an informative system that highlights different aspects and services offered by an archive.

This idea was later realized into a project which was submitted to the Archivio di Stato di Pavia but what concerns us here is to explain difficulties and propose a pattern that beyond this specific case.

Starting point is how to link the descriptions of the fonds (for example described in a finding aid, as well as in a pre-existent more complex information system) with the library's catalogue of the archive itself or with an OPAC.

In fact, I have always considered frustrating being unable to navigate through the bibliography which is supplied for each fonds, accessing directly to bibliographic records or to the lending service. Anyway – as you'll see from the model – the targets we appointed concern different aspects, not only literary works or fonds.

[4] The case of Tchaikovsky suffers obvious problem of transliteration, so we need to refine less the sample increasing K and consequently decrease minimum rate (P).

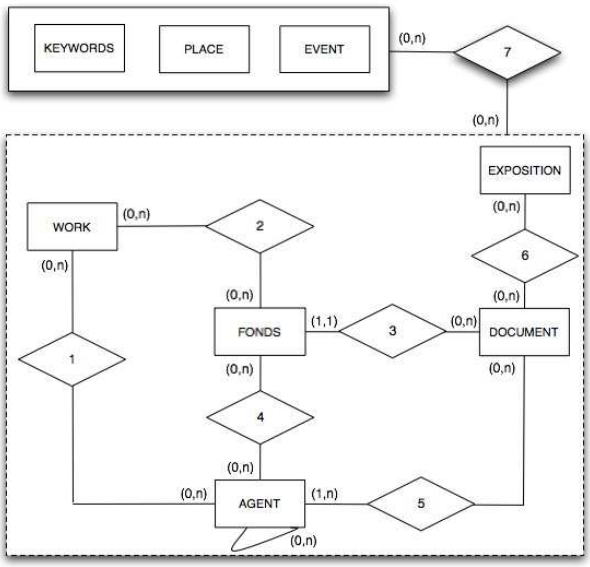

Fig. 1. Entity/Relation model. The relations are: 1- writes/is written by; 2 – is bibliography of/has as bibliography; 3 – is part of/has as part; 4 – created by/has as creator; 5 – writes/is written by; 6 - is part of/has as part; 7 - is related to/has as subject.

In this case we can identify three groups of entities: agents, objects (fonds, works, documents and exhibitions) and access points (places, events and keywords).

It's to be noticed that we provided a single entity for the agents group: this is extremely important with regard to the debate among archivists; the point – as we have often repeated - is to identify one single ontology with different descriptions and relations. This may seem a trivial conclusion, but I think that managing as a single ontology "Comune di Pavia" as creator and as custody represents a result that would make lots of archivists seethe. This could be expressed in a topic map through a single topic with different descriptions (and with the two different scopes: creator and custody).

For what concerns the "objects", the most important connection is between work and fonds (represented by the relationship "is bibliography of/has as bibliography"), whose purpose is to solve the problem of separating the fonds bibliography from the catalogue we emphasized previously. Work is a concept included in the first group of entities (work, expression, manifestation, item) of FRBR model[5][1]. We can easily map FRBR in a topic map and Alexander Siegel provided a lot of example in this sense[6]. We will create a set of PSIs to map the FRBR model, based on his researches, but we need to define PSI for all the relations between the entities of the first group and the others.

[5] Functional Requirements for Bibliographic Records.

[6] See <http://kpeer.wim.uni-koeln.de/~sigel/Projects/FRBR_and_XTM.html> in particular <http://kpeer.wim.uni-koeln.de/~sigel/Projects/FRBR/FRBR_with_SIPs.ltm> and <http://kpeer.wim.uni-koeln.de/~sigel/Projects/FRBR/FRBR_examples.ltm>.

It's worth mentioning the idea of online exhibitions[7], whose advantage is to navigate from shown documents to the fonds (or to the series, according to the description level) they belong to.

Finally, in this case there are three contextualization entities, a sort of simplification of those of the FRBR third group: concept, object, event, place. In this case the most important entity is keywords, with the aim of defining and create some research pathway to guide the inexpert user in navigating the archive.

About the implementation and the management of the topic map, several factors are to be considered:

- topics on works will be extracted from MARC[8] records. There still exist a few projects on the subject, however the cataloguing software used in this case is based on a MySql database, so the creation of a topic map can be realized with no big difficulty, either converting first MySql database into XML database and then working with a stylesheet XSL-T, or through a script querying the database to extract a topic map (the latter solution is the one we opt for at the moment);
- agents will be extracted from EAC[9] or EAG[10] documents (using a XSL-T stylesheet) and from MARC records itself;
- fonds will be extracted from descriptions realized in EAD[11] or EAG (using again XSL-T);
- some associations can be automatically created from MARC records (for what concerns the relationship author-work) or from EAD and EAG file (analysing the tags addressed to relationships between fonds and creators);
- documents are codified in TEI[12] and DALF[13] so again we can use a stylesheet to extract topics;
- exhibitions and contextualization entities will be included manually.

Quite obviously each entity will be linked to its description realized in its own standard format; in this way it will be possible to navigate directly from a fond description to its bibliography, to the single record MARC, all the way to the lending service or to the document delivery, if provided.

---

[7] About online exhibition see <http://www.archivescanada.ca/english/virtual/search.asp> <http://www.aabc.bc.ca/aabc/exhibit.html>.

[8] MAchine-Readable Cataloging see <http://www.loc.gov/marc/>.

[9] Encoded Archival Context, see also <http://www.iath.virginia.edu/eac>.

[10] Encoded Archival Guide see also <http://aer.mcu.es/sgae/jsp/censo_guia/Documentos/EAG.DTD.txt> and <http://aer.mcu.es/sgae/jsp/censo_guia/Documentos/Repertorio_de_etiquetas_EAG_Alfa_0.2.doc >.

[11] Encoded Archival Description see also <http://www.loc.gov/ead/>.

[12] Text Encoding Initiative see <http://www.tei-c.org/>.

[13] Digital Archive of Letters in Flanders see <http://www.kantl.be/ctb/project/dalf/>.

4. Managing related terms in a cultural system with a topic map

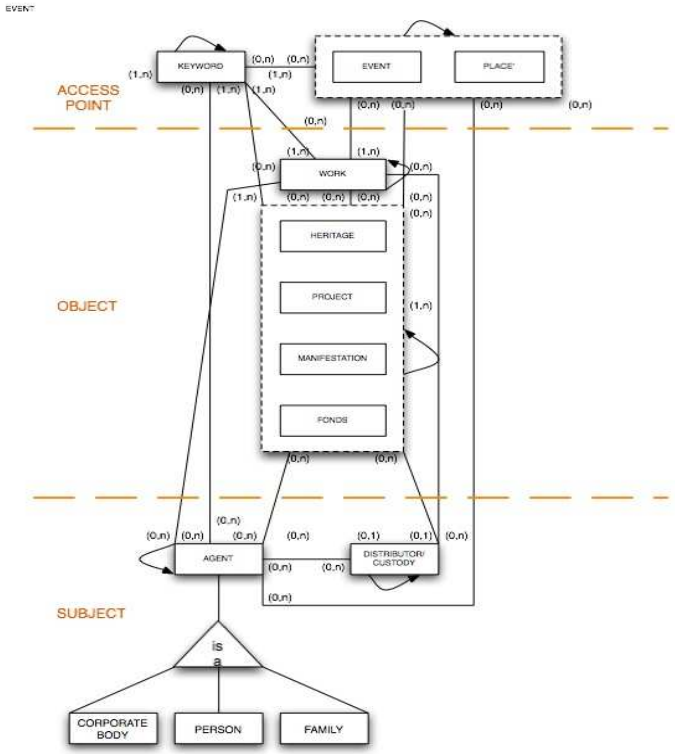

Fig. 2. Entity/Relation model of the CeDECA project

The CeDECA project, mentioned in the introduction, is a census of the cultural patrimony located in the mountain community of the province of Pavia.

In this project the principal issues deal with processing objects of a heterogeneous nature requiring different descriptive representations and different standards.

In order to develop a system that will manage the relationships between different areas of cultural heritage (for example archives, libraries and museums), it is necessary to solve various problems[2]: first of all, it is necessary to manage entities of various nature (for example, classes of objects as fonds, works, their creators, publishers, rights owners, etc.). In the case of cultural heritage repositories the challenge consists in favouring and allowing searches between analogous, though not completely overlapping, areas [3].

Another key factor towards experimenting topic maps is that the CeDECA project doesn't apply only to archival, library and museum collections, but includes a variety

of cultural resources, dynamic as well as static, such as those defined in the Minerva Project[14] [4].

The pattern this project is based on, provides three groups of entities: agent, object, access points (**fig. 2**). Regarding the agents, we chose to distinguish between custody and creator, following the well-established archivist tradition: however, in a second stage, it is possible to create on the topic map level one single ontology with different relations (associations) and different descriptions (occurrences), properly characterised through the use of scopes.

The third group of entities – access point – means contextualization entities, after the style of those of FRBR third group, we mentioned previously.

Each entity serves as the focal point for a cluster of data. The model is largely based on the principles expressed in FRBR and <in***d***ecs>[15] [5]., as well as on standards such as ISAD(G) for the multilevel description and ISAAR(CPF)[16] [6] for the treatment of creators, publishers, custodians, etc.

The analysis of attributes and relations has given evidence of many dynamic aspects related to the life cycle of an entity, to the flow of an event or even to the chronological validity of a relation. The simple use of relations and attributes defined a priori was considered inadequate because static. The need for a dynamic approach has led to consider the ABC Harmony[17] model and once again the use of topic maps. It was decided to treat the descriptions of the individual entities through a database and to manage the relations through topic maps, where topics will be automatically extracted from the database. Therefore topic maps play a twofold role, being not just subject maps, but also structure maps, through which the hierarchical complexity should be rendered.

Great affords and time were spent in developing standards aiming at enabling interoperability between archives, libraries and museums. As a matter of fact, these attempts turned out to be grids that did not entirely satisfy the requirements of either of these institutions.

We believe that 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 cohabitation of various objects. For i.e. we could have a topic “Liliana Grassi” (types: agent, creator) separated from the description level, the latter could be managed as an occurrence pointing to an EAC document, compliant ISAAR(CPF) standard.

The harmonization between different cultural heritage areas can be expressed at three levels:

- the entity level: it is necessary to produce an authority file [7] acting as the pivot between different "scopes", within different disciplinary areas, of the entity (for example a corporate body playing the role of creator, publisher, custodian, distributor, etc.). From the point of view of the description, this can be obtained in

[14] <http://www.minervaeurope.org/>.

[15] INteroperability of Data in E-Commerce Systems <http://www.indecs.org/>.

[16] International Standard Archival Authority Record for Corporate Bodies, Persons and Families.

[17] <http://metadata.net/harmony/ABCV2.htm>.

two ways: either designing a single descriptive record encompassing different[18] fields and interests, or safeguarding the specificities of every party involved and taking advantage, in a later phase, of the possibilities offered by topic maps used as a harmonization device. From the point of view of the topic map this situation will consist, either in different "scopes" or in different "topics" connected according to the degree of diversity involved in the changing role. With respect to harmonizing between variant forms of names arising from different cataloguing tradition and rules, the ADE project (Archivio Delle Entità) [8] under development in Italy, is based on the recognition of different forms, differently described, though under one single ontology. In a topic map we could have for instance the basename "Homer" scoped as AACR2 compliant together with the basename "Homerus" scoped as RICA compliant;

- the structure level [9]: one could apply descriptive models of the structure in different sectors. Particularly interesting is the application of the ISAD(G)[19][10] model, in its general rather than its specific features, to sectors different from the archives. An effort in this direction is offered, for example, by the UKOLN - RSLP[20] model. Topic maps offer the instrument to represent hierarchical relations of this type, allowing cross searches on various fields (such would be the case of a search showing on the one hand the hierarchical structure of a fonds and on the other hand the ramification of the creator connected to the latter). Moreover it is possible to deal with the single object as a monad[21] described through the appropriate specific language, but at the same time to insert it in a network providing the context. Again the works of Paul Getty can be used as the initial scheme, notwithstanding the difficulties, also of a linguistic nature, one has to cope with in order to interrelate different ontologies (it is the case of the relationship "creator – archival fonds" as opposed to the one "author - work").
- the semantic level: it is by far the level at which topic maps are used with greater profit. The difficulty, in this case, is limited to the definition of the subject terms and to their organization within a taxonomy[11], mapping whenever possible the library subject headings to the corresponding access points in archival or museum systems. In short, the aim is to supply the contextualization elements that in the librarianship field are represented by the third group entities of FRBR concept, object, event, place. We think that the realization of a semantic network, in which the objects of the speech are to be put, can't avoid confronting with these four entities.

[18] Also through a map between standards of description afferent to different words. For this purpose Paul Getty's works can be helpful for the first analysis <http://www.getty.edu/research/conducting_research/standards/intrometadata/3_crosswalks/index.html>.

[19] International Standard for Archival Description (General).

[20] Research Support Libraries Programme, see also <http://www.ukoln.ac.uk/metadata/rslp>.

[21] Meaning a single entity.

5 References

1. FRBR Functional Requirements for Bibliographic Record. 1998. <http://www.ifla.org/VII/s13/frbr/frbr.pdf>.
2. Ahmed, K.: Topic map design patterns for information architecture. XML Conference & Exposition 2003. <http://www.idealliance.org/papers/dx_xml03/papers/05-03-05/05-03-05.pdf>.
3. Simovaara, J., Vakaari, M.: Interoperability potential in data repositories of archives, libraries and museums. In: Archivi & computer. San Miniato: Archilab, 2 (2004)
4. Minerva Working Group 5: Handbook for quality in cultural web sites: improving quality for citizens. 2003. <http://www.minervaeurope.org/publications/qualitycriteria1_2draft/qualitypdf1103.pdf>.
5. Brickley, C., Hunter, J., Lagoze, C.: An Event-Aware Model for Metadata Interoperability. In: Research and Advanced Technology for Digital Libraries, 4th European Conference, ECDL 2000, Lisbon, Portugal, September 18-20, 2000, Proceedings. Lecture Notes in Computer Science 1923 (2000) 103-116.
6. ISAAR (CPF) International Standard Archival Authority Record for Corporate Bodies, Persons and Families. 2004. <http://www.icacds.org.uk/eng/isaar2ndedn-e_3_1.pdf>.
7. FRAR Functional Requirements for Authority Records. 2005. <http://www.ifla.org/VII/d4/FRANAR-Conceptual-M-Draft-e.pdf>.
8. Galeffi, A., Weston, P.G.: Il controllo d'autorità come raccordo fra sistemi descrittivi. In: Archivi & computer. San Miniato: Archilab, 2 (2004).
9. Ahmed K.: Topic maps for repository. XML Europe 2000. <http://www.gca.org/papers/xmleurope2000/pdf/s29-04.pdf>.
10. ISAD (G) General International Standard Archival Description. 2000. <http://www.icacds.org.uk/eng/ISAD(G).pdf>.
11. Garshol, L.M.: Metadata? Thesauri? Taxonomies? Topic maps! Making sense of it all. Oslo: Ontopia, 2004, <http://www.ontopia.net/topicmaps/materials/tm-vs-thesauri.html>.