Topic Maps in the eHumanities

Lutz Maicher[1], Khalil Ahmed[2], Alida Isolani[3], Aki Kivelä[4],
Sam Oh[5], Antony Pitts[6], Salvatore Vassallo[7]

[1]*Topic Maps Lab, University of Leipzig, Johannisgasse 26, 04103 Leipzig, Germany*
[2]*networked planet, Mill Street, Oxford, UK*
[3]*Signum - Scuola Normale Superiore, Via della Faggiola n. 19, 56126 Pisa, Italy*
[4]*Grip Studios Interactive Oy, Kristianinkatu 15, 00170 Helsinki, Finland*
[5]*Sungkyunkwan University LIS, Faculty Hall 40314, Seoul, Korea*
[6]*musicDNA & Royal Academy of Music, London, UK*
[7]*University of Pavia, P.za del Lino 2, Pavia, Italy*

[1]*maicher@informatik.uni-leipzig.de,* [2]*kal@networkedplanet.com,* [3]*isolani@signum.sns.it,*
[4]*akivela@gripstudios.com,* [5]*samoh21@gmail.com,* [6]*antony@cd.tp,*
[7]*vassallosalvatore@gmail.com*

Abstract

Topic Maps is a semantic technology for the integration of heterogeneous data and the convenient deployment of domain specific and semantically enriched web portals. Topic Maps is already widely in productive usage in industry and administration. We demonstrate with six projects the current usage and the opportunities of Topic Maps in the eHumanities. The projects are CeDECA from University of Pavia, RAMLine from the London Royal Academy of Music, Pansori Retrieval System from the Sungkyunkwan University Seoul, Musica migrans from the University of Leipzig, Collection of the Finnish National Gallery, and Text and Semantic from Signum in Pisa.

1. Introduction

Topic Maps is a semantic technology for the integration of heterogeneous, distributed silos of structured and unstructured data. Besides, it is a technology for the fast and convenient deployment of domain specific and semantically enriched web portals.

Topic Maps is a family of international industry standards (ISO 13250). It is in productive usage by diverse institutions in industry and administration such as the Norwegian government, the Police of Amstelland, the Norwegian post office, the Internal Revenue Service of US, and the European Space Agency. With this paper we will demonstrate with six projects the current usage and the opportunities of Topic Maps in the eHumanities.

A topic map is a highly networked semantic data store with a specific feature: within a topic map each subject of the real world is always represented by exactly one topic. A subject can be anything about information should be stored: people, places, objects, publications, genres, categories, etc. All statements about these subjects are documented by adding typed properties or typed associations to these topics. All used types are part of a domain ontology, which generally drives all Topic Maps applications [6].

Whenever two topics represent the same subject, they will automatically be merged to get one unique access point for all information about this subject. This integration model implies terminological flexibility. Generally, Topic Maps applications fit well for environments with fluid ontologies.

Within this paper we will introduce six different projects in the eHumanities based on Topic Maps. The general domains of these projects are cultural heritage, music history, and linguistics. The application focuses in the projects are heritage census, e-museum, e-learning, guided reading, semantic search, and thematic portals. Technically, the projects base on divers Topic Maps software and apply both, the integration and web portal facilities of Topic Maps. The projects are from Italy, Finland, UK, Germany and South Korea.

2. CeDECA[1]

CeDECA[2] (Centro di Documentazione Etnografica e di Cultura Appenninica) is a project for an etnographic documentation centre for the mountain area nearby Pavia (Lombardia, Italy near Milan), lead by the University of Pavia and funded by the Gruppo Azione Locale Alto Oltrepò and by Regione Lombardia. It is a Topic Maps based project and web portal for a census of the cultural patrimony located in the province of Pavia. In the case of CeDECA the covered cultural heritage contains of books, articles, pictures, music, persons, architectural heritages, archives, manufactured products, immaterial heritages, etc.

For this type of census project, the main challenge is the impossibility to define a priori the types of objects that will be described. But obviously they will emerge from the first census. Consequently a very flexible tool was needed to adapt to the growing ontology and to meet different descriptive standards' requirements [1].

Figure 1 Screenshot of part of *Giacomo Sala*'s page, one of the most famous piffero players in Italy. In the rest of the page videos, photos and audios related to Giacomo Sala can be seen.

The final ontology is organised, at top level, in seven classes (seven rooms, as in a virtual museum): the rooms of places, of museums, of folklore, of music, of history, of memories, and of books. In addition, there are two transversal rooms: the room of agents (persons, corporate bodies and families) and the room of multimedia items (image, audio, and video).

Topic Maps seemed to be a good solution to manage a fluid ontology and to meet the needs to manage different types of name (nicknames, names of places that change over the time, dialectic forms etc.) and possible variants (for example different way to write dialectic forms).

For data input, the Topic Maps editor Topincs [2] (freeware for non commercial use) was used. In wiki mode Topincs appeared to be very clear and efficient even for people who don't know the technology used.

In the frontend two graphical modes are added on top of the textual navigation: a simple navigation of the hierarchical ontology and a visualisation of the associations using Hypergraph[3], an open source java library for graphic rendering of hyperbolic trees with emphasis on Topic Maps. Lastly, Topincs allows export of the whole topic maps (or a single topic) into several Topic Maps formats such as LTM, XTM, JTM for exchange purposes.

There are some features in CeDECA that could be improved in the near future. It is aimed to work in three different directions: on research, by implementing facets refinement; on multilingual support, by providing at least an English version (from the technical point this could be easily implemented using linguistics scope notes); and finally on a general growth of object numbers (at the moment CeDECA includes 1.600 topics and 4.000 associations), by taking a census of also provinces nearby that have the same cultural historical background of the area described by CeDECA.

Speaking on this, it's quite interesting to underline how the revision phase led to an exponential growth of explicit associations: after the first data input cycle there were less than 2.000 associations on 1.200 individuals, on average they doubled in the revision phase. We foresee that new data entries and researches could make these numbers even more significant,

---

[1] Author of this part is Salvatore Vassalo.
[2] http://www.cedeca.it

[3] http://hypergraph.sourceforge.net/

highlighting the potential and simplicity in building up topic maps with different types of associations and different role types, aka a fluid ontology.

3. RAMline[4]

The RAMline[5] was an important part of a long-term research project at the Royal Academy of Music into the mapping of musical history. The goal of this musicDNA project is to provide a model for connections between people, places and musical works that can be used not only to navigate Music History, but that can also be used by individuals and organizations to chart their own musical development and link that in to the wider context of the network of influences and pathways between musical works and musical events. As a prototype system, the RAMline demonstrated many of the concepts of this model using Topic Maps.
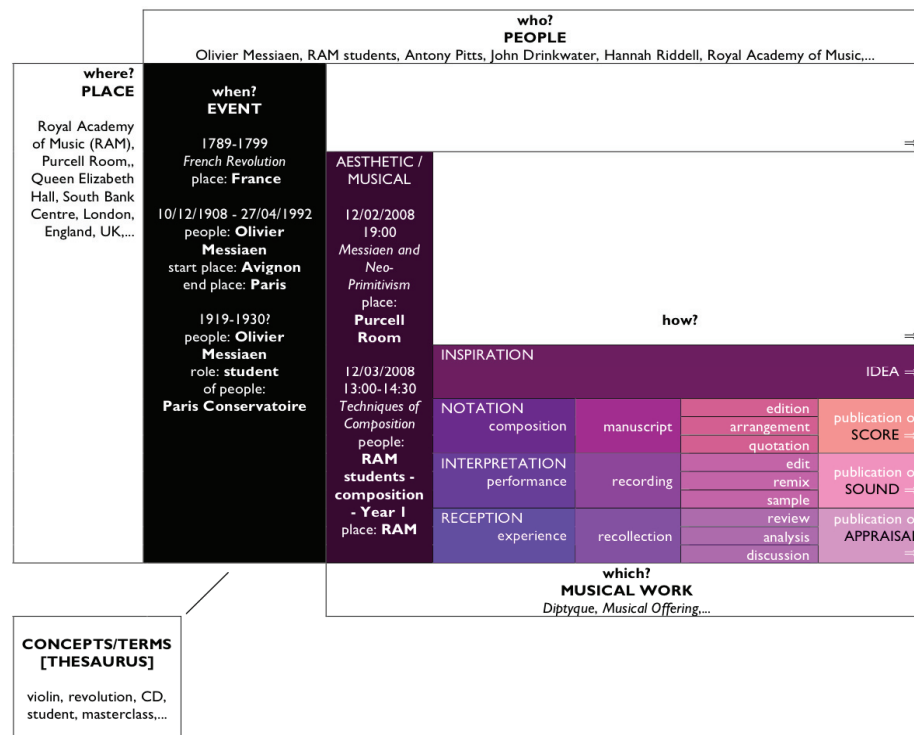
events and event participants and between the event and the location(s) where the event occurred. Occurrences are used to record the start and end dates of events, whereby fuzzy definitions are allowed. Events can also be related to other events using associations to represent cause/effect and part/whole relationships (e.g. the performance event of a particular work may be part of a concert performance event).

Aesthetic Events are a subclass of event that extends the basic event information with information about the work that is involved in the event. Musical Events are Aesthetic Events that are restricted to involving only musical works.

At the heart of the musicDNA ontology as trialled in the RAMline are the subclasses of Musical Event which are used to describe the different types of musical activity contained in a particular event. The Musical Event class has four subclasses that represent the four key domains of process in Music (inspiration-notation-interpretation-reception). The model recognizes that a musical work exists as an idea or collection of ideas in someone's head before it's written down, and then it exists as a set of instructions - the sheet music - which can then be interpreted by someone (else) and turned into actual sound, and this sound can be experienced and responded to by a third party - someone else again - at which point the work exists in the memory of its reception and consequent thoughts (often captured in print or in a broadcast – and which may form part of

Figure 2 Overwiev of the RAMline conceptualisation

The fundamental elements of the RAMline are events, where an event is a confluence of people, places and musical works that occurs within some defined (but possibly unknown) period of time. Events are modelled in the topic map as topics with associations used to record the relationship between

another, non-musical work). Within three of these four domains the same four-fold distinction between creating, capturing, connecting and communicating occurs and is modeled using further subclasses of each class that represents the domains.

At the end of the chain are the resources - scores, books, recordings, articles and so on. These resources are often the result of a communicating/publication

---

[4] Authors of this part are Khalil Ahmed and Antony Pitts.
[5] http://cd.tp/RAMline.html

type event, but they may also be involved in other kinds of events (such as when a particular instrument is used in a performance). Resources are also represented as topics, with associations used to indicate when the resource participates in or is the result of an event.

Finally a dimension is to be added to the model to enable a narrative to be layered on top of the index. Events in the RAMline may be part of a student's portfolio of studies in a conservatoire, or they might be the writings of a particular blogger.

The first Topic Maps implementation of these concepts made use of the NetworkedPlanet TMCore topic map engine[6] and an integration with Microsoft SharePoint. The RAMline interface, although somewhat restrictive, enabled non-technical users (in this case students taking an elective course in Musical History at the Royal Academy of Music) to enter the details of events, people and places using a familiar forms-based interface. Presentation of the results was achieved through a web service that exposed the relationships contained in the topic map in a simplified XML format suitable for processing in a separate web application.

Since the conclusion of the RAMline prototype phase, the research team has begun a much wider implementation of the musicDNA model – which has its first spin-off release this summer in an iPhone application called musicGPS.

4. Topic Maps-based Pansori Retrieval System (TMPRS)[7]

Pansori(판소리) is Korean folk music, a domain in which Koreans are familiar. Koreans have many opportunities to hear Korean folk music and it is studied by students during the early years of their schooling. In a first step the domain of Pansori should be examined in some detail.

The *Pansori(판소리)* is being used for two major reasons. One reason is to represent a knowledge domain for this study, the domain of Korean Folk Music, and the other reason is because it represents a specific and narrow well-structured knowledge base which has a clear meaning as a particular branch of Korean Folk Music. The *Pansori(판소리)* domain can be divided into several sub-groups. The first group offers relationships among *singers* of Pansori. Singers are associated with *genealogy(계보)*, *sect(유파)*, *tone(조)*, *part(대목)*, and *pansori(판소리)*. For example, *Sanghyun Cho(조상현)* as a singer belongs

Figure 3 TMPRS Topic Types Example

to the 'present-period' singer genealogy, to *KangSan-Je(강산제)* sect, is famous for singing well a particular song called, *ChunHyang-Ga(춘향가),* and his teacher is *Nokju Park(박녹주)*. A famous singer of *JinYang-Jo(진양조*) as a tone *is KwangRok Song(송광록)* and *Mansoon Park(박만순)* is well known for singing the *Love-Song(사랑가)*.

The second group forms relationships with Pansori. Pansori has relationships with *singer(명창)*, *part(대목)*, *sect(유파)*, *region(유명지역)*, *editorials(사설집)*, *composer(작곡가)*, *dunum(더늠)*,

Figure 4 TMPRS Topic Search Example

---

6 http://www.networkedplanet.com/Products/TMCore/
7 Author of this part is Sam Oh. The prototype of TMPRS is not online available.

body(바디), tambour(고수), and rhythms(장단). There are further complicated relationships among many groups in the Pansori knowledge structures which should not described in full detail here.

The TMPRS was implemented based on Topic Maps with an ontology derived from the domain modeling. Ontopia[8] (which is Open Source today) was used to implement an interface to demonstrate the Topic Maps based Pansori retrieval system.

Topic types are placed in the top as it shown in the figure 3 above. If a user clicks a topic type in the top menu, the interface shows its hierarchical relationships, scoped names, instances, and occurrences. If the user clicks any topics, the system displays associations in the left section, occurrences in the middle section, and its hierarchical relationship in the right section. Therefore, the user can identify all kinds of associations, and occurrences related to that topic in one display window. For example, if a user finds Chun-Hyang Ga (춘향가), which is an instance of Pansori as a topic type, s/he can see all the associations and information resources in a one page display shown in the figure 4.

To conduct a performance measurement of TMPRS, the 'Pansori.com'(CPRS) was selected because it contains exhaustive contents, well comprised categorization, and various types of data such as sounds, pictures, and text that are related to Pansori. In addition, it has been kept updated since 1996 and the system statistics reveal high traffic by users. The detailed report of this evaluation is [3].

To conclude, TMPRS performed well in complex tasks rather than in simple tasks. TMPRS' methods of modelling and structuring information allowed users to locate information with more information, and seemed to promise fewer browsing and less time. It also provided a conceptual understanding of a domain.

5. Musica migrans[9]

Within the project *Musica migrans*[10] the life courses of 19th century eastern European musicians are manually

collected from an impressive number of different archives in this region. The project is a cooperation of the Topic Maps Lab[11] and the Institute for Music History at the University of Leipzig. It is funded by the authorised representative for cultural and medial affairs of the Federal Government of Germany.

The domain of life courses is full of relationships – persons and their places of education, work and performances. The seven primary topic types used in Musica migrans are Persons, Places, Nationalities, Confessions, Professions, Sources and Dates. The individuals (or instances, from a technical perspective) are connected through typed associations.

As usual in Topic Maps portals, the information and interaction architecture is driven by this underlying ontology. There exists for each individual in the data,

Figure 5 Screenshot of the individual page in Musica migrans for *Slavka Atanasijević*, with links to the individual pages for *Osijek* and other subjects.

like the person Slavka Atanasijevic, a dedicated page in the portal, which contains all information about this person, as it is shown in the figure above. Besides the depicted HTML representation, PDF dossiers for each person are created out of the data and the life courses are drawn in maps by using the Google Maps API.

Exemplary for the complete ontology, an excerpt around the type *Person* is described here. All persons are instances of the Person-type. People are born at a specific place at a specific date. Both may be known or not and may be more or less specific and accurate and backed up by a source. Like places, also dates are first-class topics, having their own properties and associations and being presented on specific individual

pages. While the publicly available front end of Musica migrans currently allows only one birthplace and birthday, the Topic Maps backend behind would handle any number of these associations. The integration of possibly contradictory knowledge from different sources is essential for music historians. Besides the birthday and birthplace, people in Musica migrans also have a date and a place of death. Again, both may be more or less specific but formalized by their identity and associations. For the music historians, also educational periods, occupations and engagements are relevant. These periods cover a given time and take place at specific institutes in specific cities. The topic map behind the whole portal is open accessible and can be used for any third party attempt.

Technically spoken the portal is a web application based on QuaaxTM[12], a third-party, open sourced and sporadically maintained PHP topic maps engine. All data is persistently stored in a relational database and can be exchanged by Topic Maps serializations. Today the system handles 26.000 topics and 34.000 associations.

In the meanwhile the work on the second phase of the portal is ongoing. The main focus in the new project is the collection of musical institutions in this time period and area. During this project, the whole Musica migrans infrastructure will be moved to RTM [7], which is an open sourced and actively maintained topic maps engine in Ruby. For performance optimization the RTM engine is build on top of Ontopia, the currently open sourced, industry scale topic maps suite in Java. With these architectural decisions, the expected growth of Musica migrans in users and traffic can be supported.

Open issues in Musica migrans which needs further investigation are the standardization of temporal qualifications of statements in Topic Maps, the usage of a gazetteer for (historical) places, and the adoption of a standardized ontology for music history.

6. Collections of Finnish National Gallery[13]

The Finnish National Gallery (FNG) is the largest art museum organization in Finland consisting of the Ateneum Art Museum, the Museum of Contemporary Art Kiasma, the Sinebrychoff Art Museum, and the Central Art Archives. There is a constantly growing need from the audience to present all the artworks which are not exhibited in the museums halls. To answer the call, FNG started a project to design and build a web portal to promote the art collections of the

FNG. Based on positive experiences in other projects[14], it was decided to use Topic Maps and the open source Topic Maps application Wandora[15] in this project. The FNG Collections project[16] started in the beginning of 2006.
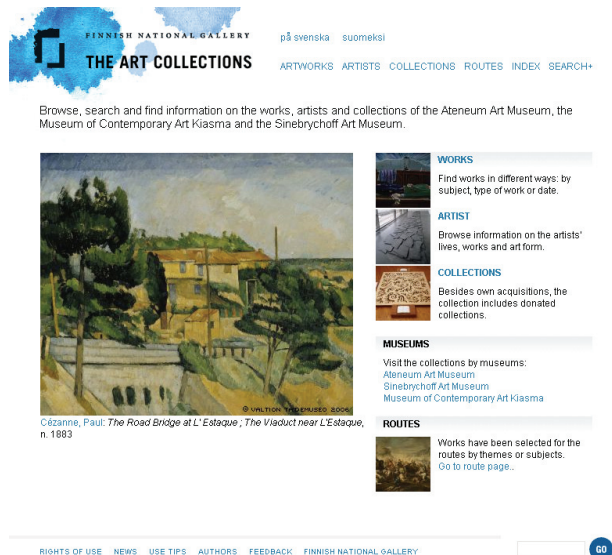

Figure 6 Screen capture of Collections of Finnish National Gallery portal reviewing narrow top navigation structure.

Generally speaking the project had three distinct steps: gathering and refining data, designing page templates, and setting up the server.

The most important part of the FNG Collections web portal is data. Most data, including artist and artwork related information, was transferred from FNG's information system Muusa[17] as a Topic Maps dump. It took several attempts to finish the transformation scheme from SQL to Topic Maps. However, when the transformation was available, Muusa data could be updated separately from other data. This feature is important as the FNG personnel continue updating data in Muusa.

Within the FNG Collections project the Muusa data was enriched with several manually generated data packets. The first data packet was the Iconclass[18] vocabulary used to annotate artworks of the Museum of Contemporary Art Kiasma. The Iconclass vocabulary was converted to Topic Maps and merged to the artwork topic map. Other FNG museums used ad hoc vocabularies developed by the museum personnel

---

[12] http://quaaxtm.sourceforge.net/
[13] Author of this part is Aki Kivelä

[14] The "Town Again" project is described at http://www.fng.fi/fng/rootnew/en/kehys/town.htm
[15] http://www.wandora.org
[16] http://collection.fng.fi
[17] http://www.muusa.net (in Finnish)
[18] http://www.iconclass.nl

during decades. To solve the distinct-vocabulary problem, a bridge connecting same keywords in both vocabularies was built.

The museum personnel write short artist and artworks related texts for exhibitions and publications. The second enriching data packet includes these curated texts. And the third data packet contains all GUI texts.

The combination of the Muusa dump and the three enriching data packets was the most critical task in the project. It was a real case for the semantic integration technology Topic Maps. As all subject identifiers in the topic maps were carefully chosen, all separate data packets could be merged easily into a seamless entity.

At the same time portal's visual look was designed by a graphic designer and static designs were converted to layout templates.

All programs created during the project were written in Java language and Tomcat was chosen as the server application framework.

The portal has a narrow top level navigation structure (see Figure above) allowing users to browse artworks, artists, and collections with different perspectives. Below this top level navigation the portal structure obscures due to rich linking. As almost everything is a link, user can continue browsing portal without ever coming back to the top level. In our experience this is very typical feature of web portals using Topic Maps and semantic technologies in general. As an implicit consequence the average ratio of browsed pages per user tends to be rather high[19].

The portal supports three natural languages: Finnish, Swedish, and English. However, it was very laborious to translate all content to three languages. As a consequence the portal is language asymmetric.

The FNG Collections portal opened July 2007. Currently the portal contains over 33.000 artworks from 3.600 artists, and over 18.000 artwork images. In

Topic Maps terms the database contains over 100.000 topics and over 500.000 associations.

7. Text and Semantic[20]

Signum[21], computer science research centre on humanities computing at Scuola Normale Superiore (Pisa), is an advanced centre for the computer-based research on digital humanities. Its mission is to define and develop new advanced tools for the humanistic domains. Signum gained a remarkable experience in analysing and processing humanistic texts, which resulted in the development of a search engine for XML documents: TauRo-core[22].
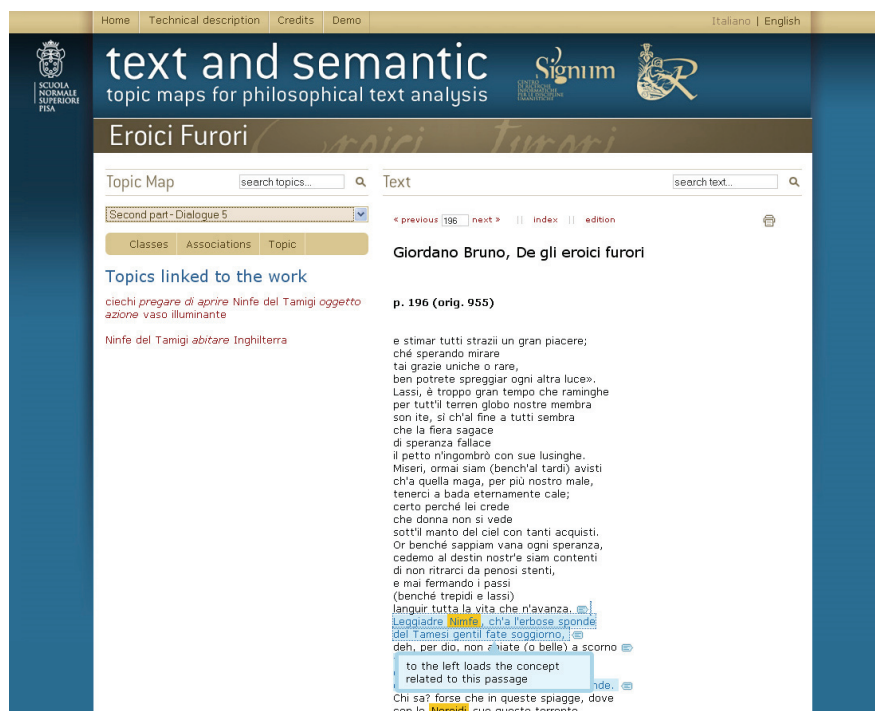

Figure 7 Screenshot of the individual page in Text and Semantic which shows the link between topic map and text.

In this research field in collaboration with the Istituto Nazionale di Studi sul Rinascimento[23], a demand has emerged for a system aimed at developing semantic research and aids for the reading and interpretation of philosophical texts. So it was decided to carry out an application to perform semantic analysis of philosophical texts in order to pass from the interpretative level to the expressive one [4] according to the targets of Dino Buzzetti's research [5].

The main goal of the project is to create a system for managing interpretations of philosophical texts by

topic maps. At present an online system was created which allows guided reading and semantic search of the 16th Century text of Giordano Bruno: *De gli eroici furori*[24]. The choice of *De gli eroici furori* was primarily determined by its subdivision into various dialogues, that permits to follow the diachronic development of concepts, thus serving as an ideal case study for the application.

For the realization of the prototype, it was decided to delineate three interpretative variants revolving around a single conceptual unit (**intellect**), and to consider three different dialogues: I,1; I,4; II,5.

In particular the concepts were described through topic maps by the XTM (XML Topic Maps) standard and also text was described using TEI (Text Encoding Initiative) standard.

Using XML in representing of the level of expression (TEI) as of the level of content (XTM) was the precondition for the search engine for XML documents TauRo-core to link both levels. In addition MySQL was used to manage interpretative variants [1].

In the prototype both text and concepts can be searched by TauRo-core: query results consist in a list of occurrences with them related context and in a list of topics relevant to the interpretation. Users can navigate separately the work and its interpretation. Besides, during the navigation users can choose to jump from topic maps to text - activating a search for topics' variants within the work - and from text to topic maps, selecting highlighted passages linked to specific concepts (see the figure 7).

As the prototype shows, the reading of philosophical texts is aided by concepts' extraction through topic maps. In addition, there is a given demonstrative example of the validity of Buzzetti's theory, according to which it is possible to pass from the level of the expression to the level of the content and vice-versa. In the future we hope to enhance this application to include new features, for example extending the system to manage and compare several texts and their related topic maps.

8. Conclusion and Outlook

Within this paper we have demonstrated the usage and opportunities of Topic Maps in ongoing eHumanitites projects. Topic Maps as semantic integration and web portal technology can be used in a broad variety of challenges the eHumanities are faced with. The main benefit of using Topic Maps is the support of environments with fluid ontologies, combined with the subject centric data organisation. The need for having

information hubs for any subjects is a common use case we observed in various eHumanity settings.

There is a vital development of new software tools and ongoing research in the Topic Maps area. Current main developments are multilingual support and localization, model driven portal generation, flexible and performing integration of heterogeneous legacy data, domain specific APIs, implementation of standardized query languages, generation of natural language output, question and answering on top of topic maps, subject centric full text search, web 3.0 support as linked data, and semantic repowering of current web tools.

We want to emphasise, that projects based on Topic Maps hence have a standardized fundament which allows to profit from all these developments on the long term.

References

[1] D.M. Zorich, G. Waibel and R. Erway: *Beyond the Silos of the LAMs: Collaboration Among Libraries, Archives and Museums*, OCLC, Dublin Ohio, 2008. Published online at: www.oclc.org/programs/reports/2008-05.pdf.

[2] R. Cerny: *Topincs Wiki - A Topic Maps Powered Wiki*, in *Scaling Topic Maps: Third international Conference on Topic Maps Research and Applications, TMRA 2007 Leipzig, Germany, October 11-12, 2007 Revised Selected Papers*, L. Maicher and L. M. Garshol, Eds, Springer-Verlag, Berlin, Heidelberg, pp. 57-65.

[3] S. Oh, O. Park: *Design and Users' Evaluation of a Topic Maps-Based Korean Folk Music Retrieval System*, in *Leveraging the Semantics of Topic Maps: TMRA 2006 Leipzig, Revised Selected Papers*, L. Maicher, L. M. Garshol and A. Sigel, Eds, Springer-Verlag, Berlin.

[4] D. Buzzetti, M. Rehbein, A. Isolani, C. Lo Rito, C. Genovesi, D. Marotta, M. Matteoli, C. Tozzini: *For a Dynamic Model of Textual Variation: What Do We Need?* In: *Digital Humanities 2009 Conference Abstracts*, University of Maryland, College Park, 2009, pp. 8-12.

[5] D. Buzzetti, *Digital Editions and Text Processing*. In: *Text Editing, Print, and the Digital World*, edited by M. Deegan and K. Sutherland, Aldershot, Ashgate, 2009, pp. 45-62.

[6] L. Maicher, B. Bock: *ActiveTM - The Factory for Domain-customized Portal Engines,* In: Journal of Universal Computer Science, 14 (Special Issue Triple-I 2008).

[7] B. Bock: *Ruby Topic Maps,* in *Scaling Topic Maps, TMRA 2007 Leipzig, Germany, October 11-12, 2007 Revised Selected Papers*, L. Maicher and L. M. Garshol, Eds, Springer-Verlag, Berlin, pp. 172-184.

---

[24] http://textandsemantic.signum.sns.it