Qualitative and Quantitative Methods in Libraries (QQML) 2: 121 – 128, 2012

An effectual approach for a data and information management for humanists

Frank Förster[1] and Bernhard Thalheim[1]

[1]Christian-Albrechts-Universität zu Kiel, Germany

Abstract: We present a generic database schema enabling different contextual points of view on a heterogeneous mass of scientific data and information within a multidisciplinary research environment. We present a method with which scientific data can be stored in a self-defined set of entities and relations, so that it becomes possible to represent the context for each research question in its own conceptual ER schema within one single relational database and without changing any involved schema. Acknowledged ontologies as particular schemata, involving bibliographic, biographic, or artifact information, can be included. The intersection of the metaphorical sphere within texts, feature specialties of material artifacts, and geographical terms allows a congruent text-artifact-map-transformation. Examples shall illustrate the broadness of our approach.

Keywords: information and knowledge services, e-science, databases, ontologies, knowledge mining

1. Introduction

An effectual approach for a data and information management for humanists has to be aligned with the researchers themselves. Humanist researchers define their aim of work by formulating questions based on texts, pictures, or artifacts, and, at the same time, by conceptualizing (more or less elaborative) philosophical questions.[1] The fluctuating intersection of the metaphorical sphere within a textual document with self-defined artifact information plus spatio-temporal relations is dependent on a particular research question. Humanist scientists tend to build up an individual database in order to conduct their semi-quantitative (and semi-qualitative) research study. It becomes important for them to create or even reuse an own corpus of texts, and to arrange any other set of entities that are intended to be analyzed through the eyes of a specific problem. Nevertheless, it is inadequately for their work to rely *only* on metadata provided by library catalogues, or archives and museum inventories, yet it is needed.

Unfortunately, it is often the case that humanist researchers install their particular databases as a basis for their studies just for their own purposes. It is unusual to share own data and information with other researchers because

[1] Förster and Thalheim (2011).

Received: 17 August 2012 / Accepted: 27 September 2012 ISSN 2241 1925

differences between individual research questions and approaches require dissimilar databases. Singular problems involve an own methodology, or the foundation of a specific approach. Mostly, having recent humanist doctoral projects on a watch, the databases are of simple structure and not very sophisticated, although aiming at answers for a certain question. The researchers tend to abstain from relations to relevant sources and put all information into a single database sheet. If they use library, museum or archive information, they have a tendency rather to copy metadata information from relevant catalogues (if applicable) than to enrich their data by including references to standardized vocabularies. The insertion of heterogeneous metadata from the indexing practice of libraries, museum documentation and finding aids from archives into a comprehensive database is one aspect to be solved. Mainly, catalogues still remain repeatedly only information supplier in terms of completing a certain corpus. The most probable outcome of humanist research is monographs, or volumes with single-authored papers, having no need to provide a database foundation as with geoscientific data for example. Humanist scientists keep their findings for future publications within home-made and long-lasting archives.

So-called "Virtual Research Environments" (VRE) are appropriate for a particular discipline. VREs for large research communities are asked to be enhanced as generic infrastructure with modular and flexible services according to recent studies.[2] In this respect the focus of those achievements lies on reference management, on instruments which support publication of data, and on applications for an unconstrained and fast communication (e.g. wiki, blog). But that is not intended with the present study.

Our approach attracts notice to the researchers' databases containing research data and information which is relevant for data processing. The analysis of texts or artifacts results in an assorted schema of particular features. Their assemblage is driven by individual requirements. It is important for an inter- and multidisciplinary research environment to tackle the problem of data storage and processing from a more abstract point of view, and also from the beginning[3]. This is the initial position for a data and information management for humanist projects within the Graduate School "Human Development in Landscapes" (Kiel University, Germany) which is tied together by a research question crossing traditional borders of disciplines, claimed as "to detail the interactions between mankind and both its physical and perceived environment."[4]

The paper is organized as follows: In section 2 we detail specific goals, methods, functionalities, and modules of the humanist database. In section 3 we handle exemplary research projects from our related doctoral programs to demonstrate the capabilities of the planned research data and information

[2] Horstmann et al. (2011, p. 355).
[3] Fleischer and Jannaschk (2011).
[4] http://www.uni-kiel.de/landscapes/index3.shtml

infrastructure. Achievements are shown in section 4 together with a conclusion for possible future lines of research.

2. Goals, methods, and functionalities

Central aim is the development of a database concept which supports a data and information management in the humanities in an efficient way. Our approach comprises a Service Oriented Architecture (SOA) with certain web services integrated within a generic database application. Those web services are on the one hand tools for georeferenced data, i.e. gazetteers, visualization services and (2D-, 3D-)map creators (Web Coordinate Transformation Service, Web Feature Service, Web Coverage Service, Web Terrain Service, and Web Map Service). On the other hand we develop a structured meta-information system for including citation and bibliographical information, artifact information, and statistics. For this reason, singular research projects shall borrow features from metadata schemata of various proveniences. Connections shall be made for information about persons (by accessing the Virtual International Authority File[5]), about geographical terms (appropriate tools will be developed, according to Nolde et al. 2010), and about bibliographical items (e.g. WorldCat[6]). All research database information can thus be enriched by supplementary information provided via access points to intellectual or artistic contents within the digitized bibliographic universe.

The configuration of those features (in respect to the specific research question) is singular in any case. And as the focus or even one aspect of the research question changes, the database element set should change as well.[7] Our generic database solution will be developed as generalization of the CIDOC CRM[8] specification.[9] It will be realized as one simple relational database which enables singular researchers to create particular conceptual entity-relationship (ER) schemata out of the demands of CIDOC CRM and also own needs.

Database structures used in humanistic research are typically simple. The database tables use a small number of attributes for property description. These tables are interlinked in order to associate data with each other. The tables themselves are populated by a small number of rows or objects. At the same time the research portfolio drives the utilization of many different tables. The research portfolio evolves over time. The interest of the researchers in their data changes, too. This behavior results in a large set of different tables with small data sets. Thus, database applications become quickly unmanageable, difficult to use, deploy, and query. For that reason, complex legacy schemata are created, but their maintenance is often heavily delayed.

---

[5] http://viaf.org/
[6] www.worldcat.org/
[7] Terwilliger et al. (2010) presents an analogue example for object-relational mappings.
[8] The Conceptual Reference Model (CRM, ISO 2006:21127) has been developed by the Comité international pour la documentation (CIDOC), www.cidoc-crm.org.
[9] Jannaschk et al. (2011) and Bienemann (2008).

We developed a simple approach for storing such data. All data is maintained in a common repository that accepts data rows as triples consisting of a unique row identifier, a pair representing the structuring (relation name and attribute name) and a value that is used for that attribute within this row within the given table. This structure is called generic structure. Therefore the row (4711, Name="Thalheim", Institute="CS", Phone="4472") in the table PERSON is represented by three triples: (4711, (PERSON,Name), Thalheim), (4711, (PERSON, Institute),CS), and (4711, (PERSON,Phone), 4472). These triples contain the same information as the row. At the same time they can be stored in a simpler fashion. We might use an equivalent approach for storage of associations among data objects and might also enhance the triples with additional data. For instance, the mentioned row – when inserted by "Förster" – is represented by (4711, (LOG_INSERT,Issuer), Förster). Therefore, our approach is also capable to store metadata, i.e. the data about data.

Whenever a researcher wants to add a new table, then the profile of the table is used for the introduction of two special generator functions:

- a storage function allows to generate the triple mapping for each object;
- a retrieval function allows to reconstruct the original object from the triples.

These two functions are supported by any database management system since they can be specified as views defined over the database structure. The classical approach for an evolution of database structures results in a revision of the entire database system, in recoding and rearranging the storage engine. If an application is characterized by eager changes whenever necessary, then classical database technology cannot handle it.

Our approach is based on a separation of concerns (SoC). We separate storage from deployment of data. Classical database research combines both. This combination was mainly caused by the limitations of technology until 2000. Nowadays, machines are faster and storage facilities are far more advanced. The SoC may also be supported by two special machines: a *generic storage machine* which uses our approach and a *computation machine* that orients on high performance computation and is fed by the storage machine. Both machines may be decoupled. It seems, however, that such separation is limiting the user in deploying data in a sophisticated form. Data are often not only retrieved and selected row-by-row. Users want rather to see their data in a changed structuring, with some specific display and based on some computation. If we would use the approach presented so far, then we would have to transfer such requests to the computation machine. This would lead to a heavy burden for the programmer.

However, we can provide another solution.[10] Actions that are bundled together can be represented as short stories within the application, so-called *mini-stories*.

[10] Similar solutions have already been developed for web application systems (see Schewe and Thalheim 2007) and non-relational data stores (Amazon SimpleDB 2009).

They are represented by graphs that connect simple steps. A simple step is either a storage step or a retrieval step. Therefore, we may now use such a mini-story for a generation of a transaction that operates on the storage machine. The mini-stories may be combined to larger ones, e.g. research stories, characterizing the methods of a humanist scientist.

We can now represent mini-stories and research stories of users within the user space. Each user has a specific profile consisting of the data structures addressed by the researcher and of the stories employed by a user. The user space is assembled whenever a user initially uses the system. It can also be modified whenever there is need for it. Furthermore, we may assign *roles* to users. These roles are associated to some of the mini-stories. Therefore, a user can choose (together with the login into the system) the current profile within the available profile so far. An example of a mini-story is a search function. It turns out that there are seven different kinds of searching ranging from zapping and browsing to property-value-driven targeted search.[11] Therefore, we must specify first which search function is meant. Depending on that search function we choose a mini-story, supported by corresponding web interfaces. The same approach can also be used for insertion and modification of complex data. We may use a pattern approach for such functions.[12] We observed a small number of insertion patterns and a small number of modification patterns. These patterns can be combined to mini-stories and thus be supported by the generic storage machine.

It seems that the initialization becomes a bottleneck within our solution. This is however not the case. A user may select the mini-story which fits best to the needs. This mini-story is then taken to the user space and becomes an element of the user profile. Whenever a user wants to change the profile, then we might use the same approach to support that. If a user does not find an appropriate mini-story, then we start an extension of the generic facilities.

Our approach has a number of advantages. We can very flexibly react on new requests. We may support any data structuring, mini-stories and research stories. In the classical setting such tasks result in programming tasks which often require support by specialists since humanist researchers might not have deep skills in programming languages such as SQL. In the SQL setting the user may write the query on fly. We prefer a good support for each user.

We may also collect all tasks in a repository which users want to perform with the system. This repository is a collection of generic functions, mini-stories, and research stories within a community of practice. The course of actions would be as follows: A user requests a specific function. This function is mapped to the support facilities. All other users may now want to add these support functions to their profile after the derivation of the new support facilities. Our approach allows therefore the development of *community-specific* procedures. As a result,

[11] Düsterhoft and Thalheim (2004).
[12] Feyer et al. (1998).

we may use this approach for the creation of a community workspace. This workspace is continuously evolving as long as the community is active.

We are currently developing an approach to deliver this system as *open source solution*. In this case any other group of researchers may download the system, modify it according to the rules of open source communities and use it for free.

The limitation of our approach is also obvious. We do not target applications that must handle "big data" in a very efficient way. We also do not target high performance web services. Instead, we want to be evolution-robust and allow the handling of very complex schemata as well.

3. Example Projects

The current section provides details about exemplary projects pursued within the framework of the Graduate School "Human Development in Landscapes". The researchers challenge archaeological and natural science data with a literature survey in order to conduct philosophical elaborations enriched by a quantitative basis. This multitude of approaches fructify a particular research question from different scientific viewpoints, dealing with the perception of a specific geographic area (*landscape*) from prose texts or research literature in conjunction with bare facts derived from corings, measurements, or excavations. The data and information management supports the merging of both approaches.

The first exemplary research project is dealing with animal dissemination across the Mediterranean in Ancient times. Ultimate goal is a map which shows the movements of animals as result of economic transactions across the area. The material comprises excavated animal bones, and relevant site reports about it. The taxonomy of animal species and osteometric data of skeletal elements is of high importance. Conclusions about certain subsistence strategies or ritualistic acts can be drawn with combining individual provenience information in contrast to find places while considering archaeological features and age determination. The humanities database allows linking both features from faunal assemblages with findings in ancient literature. The vagueness of information in literary resources (spatial and temporal context) can be substantiated with more precise data from excavations and measurements.

The second exemplary research project sheds light on the interpretation of villa images from Roman times. Ancient villa mosaics or paintings do not only depict the house itself, but also trees, animals, gardens, and other landscape elements. But it is important to know that they stay an artistic creation and thus are always evidence of an imaginative act. That means that there is a high variability in whether those villa images show a realistic copy of an actual villa or a dream-like, ideal composition. A formal description regarding standardized typologies is foundational for the respective database which is filled with the combination of all pictorial elements plus information from literary accounts, and amended by data from pollen or soil analyses from the surface of the possible place of the depicted villa. It results in conclusions about architectural developments, distribution of flora and fauna, and "concepts of life".

The third exemplary research project heads for a map of the island of Menorca providing information for all archaeological find spots from excavation reports and literature sources. The respective statistical database comprises a self-defined set of features in order to allow empirical generalizations about certain prehistoric monuments in conjunction with their demographic, ecological, economical, and geographical embeddings.

The fourth exemplary research project tries to reconstruct the environmental history of the Peloponnesus.[13] The region has been famous for the labors of Heracles in ancient Greek times with the outcome of a diversity of written records of both poetic and prose texts. Poetic texts refer mainly to the mythological history. Although landscape descriptions are included here to a certain degree, they are more vividly present in prose texts. Several lakes, like Lake Stymphalos (Λίμνη Στυμφαλία), provide an ideal situation of detecting the landscape history by retrieving sediment cores. Particular events (e.g. volcano eruptions) are still present in the cores and can be linked to relevant descriptions in literature. The database infrastructure allows a simple combination of both aspects by interconnecting particular spatio-temporal features.

The fifth exemplary research project is doing a bibliographic survey of the publication history of the English author Joseph Conrad (1857–1924).[14] The inclusion of FRBR elements into the data model is of high significance here, and a spatio-temporal presentation helps in identifying and tracing certain lines of thoughts and developments within the reception history.

4. Achievements and Conclusions

To support small-scale projects of humanist researchers effectually, it is needed to consider their specific methods of working. As they often conduct long-lasting research processes it seems valuable to maintain evolving database schemata. As they often use databases of simple structure it seems appropriate to provide a suitable infrastructure. And as singular steps of storing and retrieving data are similar, they can be split up into adequate mini-stories; and several mini-stories can be coupled with each other to gain equivalent research stories. The scientists develop their databases in close cooperation with a programmer.

It is also aim to pull together initiatives from libraries, museums and archives with humanist activities having a university background to gain a multiple perspective of research. Our data and information management acts as a mediator between both spheres, enhancing unique research questions from both viewpoints. The appropriate integration of standardized information (geospatial, bibliographical etc.) due to its implicit need has to be guaranteed.

Our approach still supports individual humanist research, but sets the individuality free in transferring it into a shared research environment embedded

[13] Unkel et al. (2011).
[14] Förster (2007).

within the achievements of scientists having similar interests and also within the efforts of cultural heritage institutions.

References

Amazon SimpleDB (2009). Developer Guide (API version 2009-04-15). docs.amazonwebservices.com/AmazonSimpleDB/latest/DeveloperGuide/

Bienemann, A., (2008). Context-Driven Generation of Specifications for Interactive Information Systems. PhD thesis, CAU Kiel, Dept. of Computer Science.

Düsterhöft, A. and Thalheim, B. (2004). Linguistic based search facilities in snowflake-like database schemes. *Data and Knowledge Engineering*, Vol. 48, 177 – 198.

Feyer, T., Schewe, K.-D. and Thalheim, B. (1998). Conceptual design and development of information services. In: Proc. ER'98, LNCS 1507, Berlin, Springer, 7 – 20.

Fleischer, D. and Jannaschk, K., (2011). A path to filled archives. *Nature Geoscience*, Vol. 4, 575 – 576.

Förster, F., (2007). *Die literarische Rezeption Joseph Conrads im deutschsprachigen Raum*. 2nd ed. Universitätsverlag, Leipzig.

Förster, F. and Thalheim, B., (2011). Data and Information Management for Humanist Researchers (Text and Things). *Conference Proceedings: Supporting Digital Humanities 2011, University of Copenhagen, Nov 17-18, 2011.*

Horstmann, W., Kronenberg, H. and Neubauer, K. W., (2011). Vernetzte Wissenschaft. Effektive Forschung mit neuen Werkzeugen. *B.I.T.online*, Vol. 14, No. 4, 354 – 362.

Jannaschk, K., Rathje, C. A., Thalheim, B. and Förster, F., (2011). A Generic Database Schema for CIDOC-CRM Data Management. In: Eder, J., Bielikova, M. and Tjoa, A M., (Eds.): *CEUR Workshop Proceedings II of the 15th East-European Conference on Advances in Databases and Information Systems (ADBIS 2011)*. Berlin, Heidelberg: Springer, 127 – 136.

Nolde, M., Duttmann, R., Blaschek, M. and Klein, U., (2010). Geodateninfrastrukturen und ihre Anwendungen in der Praxis. *PIK – Praxis der Informationsverarbeitung und Kommunikation*, Vol. 33, No. 4, 244 – 251.

Schewe, K.-D. and Thalheim, B. (2007). Pragmatics of storyboarding for web information systems: Usage analysis. *Int. Journal Web and Grid Services*, Vol. 3, No. 2, 128 – 169.

Terwilliger, J.F., Bernstein, P.A., Unnithan, A., (2010). In: *Conceptual Modeling – ER 2010*, LNCS 6412. Berlin, Heidelberg: Springer, 146 – 159.

Unkel, I., Heymann, C., Nelle, O. and Zagana, H., (2011). Climatic influence on Lake Stymphalia during the last 15 000 years. In: Lambrakis, N., Stournaras, G. and Katsanou, K., (Eds.), *Advances in the Research of Aquatic Environment*, Vol. 1, 75 – 82.