# A data integration concept for an interdisciplinary research database

*Development of an archaeological and palaeoenvironmental data model for integrating heterogeneous spatio-temporal data*

Christian WILLMES [a,1] and Georg BARETH [a]
[a] *Institute of Geography, University of Cologne*

**Abstract.**
This paper presents an overview of the current state of development of an archaeological and a palaeoenvironmental data model for an interdisciplinary research database. The models are constructed iteratively by integrating heterogeneous data and adjusting the model where necessary. The integration concept is an iterative approach which combines several techniques for data model development, including semantic and syntactic integration and alignment, as well as semantic data linkage with external knowledgebases and models. The goal is to provide integrated spatio-temporal access to an existing wealth of data to facilitate research on the integrated data basis.

**Keywords.** Data Integration, Spatio-temporal, Semantic Web, Graph Data Model.

## Introduction

The Collaborative Research Centre 806 (CRC806)[2] is an interdisciplinary research project with more than 100 researchers from the disciplines of archaeology, the geo-sciences and cultural sciences, funded by the German Research Foundation (DFG)[3]. A central research database, the CRC806-Database[4], is currently under development and sets out to accomplish two main goals: The first of these goals is to provide a long-term archive and publication platform for results produced by CRC806 researchers. This aspect implements the data management policy that is mandatory for DFG-funded CRCs [7], and will be, from its data management perspective, comparable to other DFG-funded CRC research databases, e.g. [5]. The second of the two goals is to provide an integrated data basis to facilitate the research within CRC806. This paper will focus on the development of this second aspect.

Generally speaking, there is a wealth of information and data already available for the two data domains that are considered in this task. However, both archaeologists and

[1]Corresponding Author: Christian Willmes; E-mail: c.willmes@uni-koeln.de
[2]`http://www.sfb806.de`
[3]`http://www.dfg.de`
[4]`http://crc806db.uni-koeln.de` (launch of web portal in summer 2012)

palaeoenvironmentalists use a vast and ever-changing array of recording systems, all based on diverse theoretical perspectives, typologies, nomenclatures and methods [9,10]. Custom and non- (or poorly) documented data formats, and general access constraints to potentially interesting datasets, add a further dimension to the problem. The presented work in progress attempts to resolve this problem (at least partially) by integrating heterogeneous data and providing well-defined semantics to facilitate research on the integrated data basis.

## 1. Methods and technology

Both, archaeological and palaeoenvironmental, data models represent intrinsically spatio-temporal data, a factor which has been at the centre of our considerations from the very outset of data model development. As such, the model is designed in a way that it is able to undertake spatial and temporal filtering of each data record of the integrated data basis. Existing vocabularies are facilitated where possible, for example the semantics for spatial referencing are formulated using the W3C Basic Geo (lat/long) Vocabulary[5], and the sematics for bibliographical references are formulated using the BibTeX namespace of the MIT Simile project[6].

In the development of the model, an iterative bottom-up approach is applied. This means that we are developing the model from the semantics introduced by each of the integrated datasets. Accordingly, semantic entities which are not covered by the data model will be added to the model in the course of their integration. The process of semantic alignment of new data within the existing model is realized where necessary in consultation with domain experts from CRC806. The top-down approach, integrating each dataset into a semantically well-defined but static and complex existing model was abandoned, it proving too rigid, i.e. not adequately flexible in the integration process. None the less, it is possible to map the resulting data models to existing models in the two domains, that provide semantical interference.

The approach presented here is a dynamic concept for data model development; with every new dataset that is integrated into the database, the semantics of the model can be extended. By building the model using *semantic technology* [1,11], in particular by employing a graph data model [4], the development process can be dynamic and, most importantly, extension of the semantics of the models does not affect the application level [11].

## 2. Data Models

### *2.1. Archaeological data model*

#### *2.1.1. Integrated Data*

For the development of the archaeological data model, we have integrated the datasets listed in Table 1. The data stem from published databases [3] [13][7] and [12], and from

[5]`http://www.w3.org/2003/01/geo/`
[6]`http://simile.mit.edu/2006/11/bibtex`
[7]From CalPal only the Europe database intgrated so far.

project internal collections of data. All datasets were provided in tabular form, each with custom semantics and schema.

**Table 1.** Key numbers describing the integrated archaeological databases, $T$ = temporal extent (oldest and youngest artefact) in kBP (kilo years before present) of database.

| Database | Artefacts | Sites | $T$ (kBP) | Spatial |
|---|---|---|---|---|
| NESPOS | 0 | 296 | 120 - 10 | World |
| CalPal | 16897 | 4234 | 52 - 0 | Europe |
| Stage3 | 1897 | 412 | 108 - 0 | Europe |
| Project-Internal | 486 | 283 | 60 - 3 | World |
| CRC806-DB | 19280 | 5225 | 120 - 0 | World |

### *2.1.2. State of the archaeological model*

The model (see Figure 1) comprises three main objects: *Artefacts*, *Sites* and *SiteAttribution*. Most of the archaelogical datasets so far integrated into the database and its underlying model are based on records relating to artefacts. *Artefacts* are located by a reference

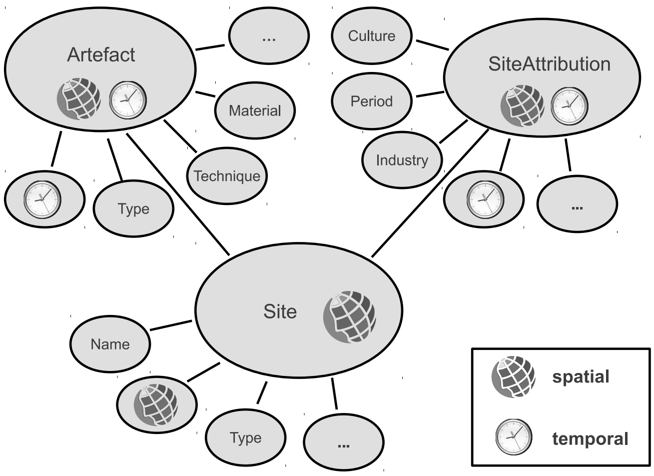


**Figure 1.** Generalized graph representation of the archaeological data model.

to the excavation *site* at which they were found. Additionally, some datasets are based on records per excavation *site*. This kind of record deals with abbreviated variables, which in most cases are derived from the artefacts found at a given *site*. Such variables are strongly connected to artefact characteristics (age, cultural attribution, etc.), but the actual reference to artefacts is not always given in these cases. For these kind of records, the object *SiteAttribution* was developed. This has the added ability that it can characterize a *site* object with additional, not generally applicable (for example only valid for a given point in time or a time range) semantics given by the site object, and thus enable site based analysis.

### 2.2. Palaeoenvironmental data model

#### 2.2.1. Integrated Data

For the development of the palaeoenvironmental data model, datasets from Stage3 [12], BIOME [6], Africa6kLSC [8] and from PMIP II [2] (see Table 2) have been integrated so far.

**Table 2.** Key numbers describing the integrated palaeoenvironmetal databases, $n$ = derived spatial datasets. $V$ = Environmental variables, $T$ = temporal extent (oldest to youngest) in kBP. $\delta T$ = Temporal periods. Spatial = Spatial extent.

| **Database** | $n$ | $V$ | $T$ **(kBP)** | $\delta T$ | **Spatial** |
|---|---|---|---|---|---|
| Stage3 | 2335 | 46 | 60-0 | 5 | Europe |
| BIOME | 6 | 2 | 18-0 | 3 | World |
| Africa6kLSC | 6 | 6 | 6 | 1 | North Africa |
| PMIP II | 7326 | 82 | 21 - 0 | 3 | World |
| CRC806-DB | 9673 | 136 | 60 - 0 | 6 | World |

The number $n$ derived spatial datasets, is resulting from the sum of the *temporal steps* per environmental variable $V$ times the number of temporal periods $\delta T$ of the database. Possible *temporal steps* are: 1 = *annual*, 12 = *monthly*, 13 = *Plant functional Type (pft)*, or 24 = *hourly values* for most cases.

#### 2.2.2. State of the palaeoenvironmental model

Each dataset (see Figure 2) is spatio-temporally referenced within the model, with a *temporal extent* describing a time range or period (with a defined start and end date and temporal resolution), or a *temporal location* describing a specific point in time, and with an *geographic extent* (a bounding box) or a *geographic location* (a point). Furthermore, the content and dataset type is classified in the semantics of the dataset objects. With this information, datasets can be filtered and accessed in spatio-temporal alignment with the overall semantics of the datamodel. The spatial data is stored internally in a processed (GIS) dataformat as well as in the original data format.

## 3. Implementation

The data integration process is not fully automated, because for each new dataset that is to be integrated, a custom translation is developed. During this development process, the semantic entities of the datasets considered are manually aligned with the current internal model by formulation of the semantic mapping in program code. If some semantic entity is not yet represented or not alignable with the current model, the entity will be added to the internal model, wich results in an expansion of the former internal semantic model.

The central RDF store containing the integrated data basis is queryable from a SPARQL endpoint. Additionaly the derived GIS datasets are accesible via an OGC standards based SDI providing WMS, WFS, WCS and CSW interfaces.

The CRC806-Database web portal[4] provides the main user interface to the integrated data basis. The web portal is a Typo3 based webapplication providing i.) a browseable

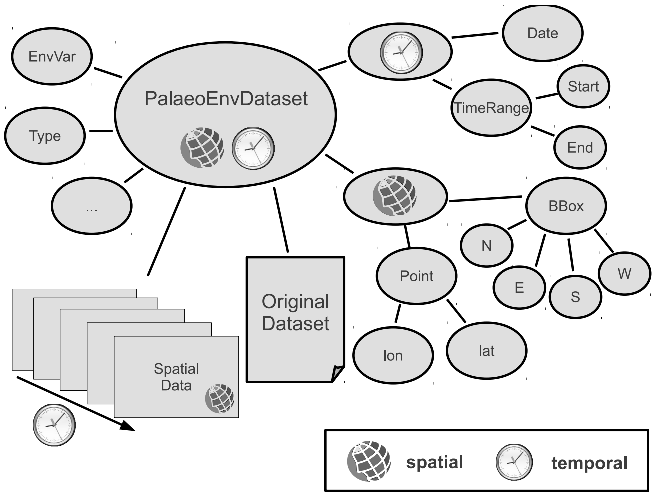


**Figure 2.** Generalized graph representation of the palaeoenvironmental data model and also simplified versions of the temporal and spatial models.

and sortable catalog, ii.) a keyword search, iii.) an Exhibit[8] based interface for faceted browsing and interactive timeline visualization, and iv.) a GeoExt[9] based WebGIS for intuitive access to the SDI interfaces. The catalog and search interfaces are implemented using the *Typo3 semantic web extension*, which allows to build user interfaces to formulate SPARQL queries addressing the central RDF store of the integrated database and rendering the results of the queries from within the web application interface.

## 4. Conclusion and Outlook

So far ~10,000 palaeoenvironmental (Tab.2) and ~20,000 archaeological (Tab.1) data records have been integrated. These are now available for integrated analyses. The rather simple but straight-forward approach to data integration presented in this paper was purposely chosen. Originally, a top-down approach was considered which would have integrated the given data into existing models. However, this approach was abandoned, not least due to its limited flexibility and its susceptibility to error. This led us to the adoption of the presented iterative approach. This approach has several advantages, such as flexibility and extendibility, though the key advantage is that the resulting data model always suits our system as it adapts organically to the demands of the CRC806-Database system and to the semantics of additionally integrated datasets. The integrated data model can be mapped easily into existing external models, and linked to suitable vocabularies and ontologies to strengthen its semantic interoperability. The main result of the work presented here is the integrated data basis and the data model derived from the integration process. Further datasets will be integrated into the CRC806 Database in the future. Project participants and database users can suggest or provide new datasets for integration. The development of ontologies for both data models, their documentation and pub-

[8] `http://www.simile-widgets.org/exhibit/`
[9] `http://www.geoext.org/`

lication constitutes a component of the PhD dissertation of one of the authors[1], and will be carried out during this project.


## Acknowledgements

The authors would like to thank the creators of the integrated databases for publication of their data and thus providing the data to the here presented project. Furthermore we would like to thank Lee Clare for improving the English in this paper, and the DFG for funding the CRC806.


## References


[1] D. Allemang and J.A. Hendler. *Semantic web for the working ontologist: modeling in RDF, RDFS and OWL.* Morgan Kaufmann Publishers/Elsevier, 2008.

[2] P. Braconnot, B. Otto-Bliesner, S. Harrison, S. Joussaume, J.-Y. Peterschmitt, A. Abe-Ouchi, M. Crucifix, E. Driesschaert, Th. Fichefet, C. D. Hewitt, M. Kageyama, A. Kitoh, A. Lan, M.-F. Loutre, O. Marti, U. Merkel, G. Ramstein, P. Valdes, S. L.Weber, Y. Yu, and Y. Zhao. Results of PMIP2 coupled simulations of the Mid-Holocene and Last Glacial Maximum - Part 1: experiments and large-scale features. *Climate of the Past*, 3(2):261–277, 2007.

[3] M. Bradtmöller, A. Pastoors, A. Slizewski, and G.-C. Weniger. NESPOS- A digital archive and platform for Pleistocene archaeology. In C. Curdt and G. Bareth, editors, *Proceedings of the data management workshop*, volume 90 of *Kölner Geographische Arbeiten*, pages 13–18. University of Cologne, 2010.

[4] J. J. Carroll and G. Klyne. Resource Description Framework (RDF): Concepts and Abstract Syntax. Technical report, W3C, February 2004.

[5] C. Curdt, D. Hoffmeister, C. Jekel, K. Udelhoven, G. Waldhoff, and G. Bareth. Implementation of a centralized data management system for the CRC Transregio 32 "Patterns in soil-vegetation-atmosphere-systems". In C. Curdt and G. Bareth, editors, *Proceedings of the Data Management Workshop, 29.-30.10.2010*, volume 90 of *Kölner Geographische Arbeiten*, pages 27–33. University of Cologne, 2010.

[6] M. E. Edwards, P. M. Anderson, L. B. Brubaker, T. A. Ager, A. A. Andreev, N. H. Bigelow, L. C. Cwynar, W. R. Eisner, S. P. Harrison, F.-S. Hu, D. Jolly, A. V. Lozhkin, G. M. MacDonald, C. J. Mock, J. C. Ritchie, A. V. Sher, R. W. Spear, J. W. Williams, and G. Yu. Pollen-based biomes for Beringia 18,000, 6000 and 0 14C yr BP. *Journal of Biogeography*, 27(3):521–554, 2000.

[7] E. Effertz. The funders perspective: Data management in coordinated programmes of the German Research Foundation (DFG). In C. Curdt and G. Bareth, editors, *Proceedings of the Data Management Workshop, 29.–30.10.2010*, volume 90 of *Kölner Geographische Arbeiten*, pages 35–38. University of Cologne, 2010.

[8] P. Hoelzmann, D. Jolly, S.P. Harrison, F. Laarif, R. Bonnefille, and H.-J. Pachur. Mid-Holocene land-surface conditions in northern Africa and the Arabian Peninsula: A data set for the analysis of biogeophysical feedbacks in the climate system. *Global Biogeochem. Cycles*, 12(1):35–52, 1998.

[9] L. Isaksen, K. Martinez, N. Gibbins, G. Earl, and S. Keay. Linking Archaeological Data. In *Computer Applications and Quantitative Methods in Archaeology conference*. CAA, 2009.

[10] E. Kansa. A community approach to data integration: Authorship and building meaningful links across diverse archaeological data sets. *Geosphere*, 1(2):97–109, 2005.

[11] T. Segaran, C. Evans, and J. Taylor. *Programming the Semantic Web*. O'Reilly Series. O'Reilly, 2009.

[12] T. van Andel and W. Davies. *Neanderthals and modern humans in the European landscape during the last glaciation: archaeological results of the Stage 3 Project*. McDonald Institute Archaeological Research monographs, Cambridge, UK, 2003.

[13] B. Weninger, K. Edinborough, M. Bradtmöller, M. Collard, P. Crombe, U. Danzeglocke, D. Holst, O. Jöris, M. Niekus, S. Shennan, and R. Schulting. A Radiocarbon Database for the Mesolithic and Early Neolithic in Northwest Europe. In P. Cromb, M. Van Strydonck, J. Sergant, M. Boudin, and M. Bats, editors, *Chronology and evolution within the Mesolithic of North-West Europe*, pages 143–176. Brussels, 2010.