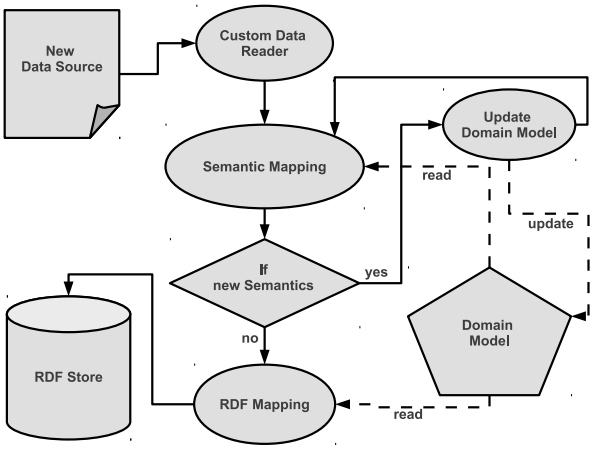

Figure 1: Flow diagram of the data integration process.

lost and that the integrated data always suits further application of analysis, models and visualizations.

A problem of the presented approach at this stage is the identification of objects that refer to the same object in reality, but are represented with different Uniform Resource Identifiers (URIs) in the integrated database. This occurs if an identifying individual/entity, for example a site name, is spelt differently or is represented in another language in the source data sets, which then results in different URIs for those objects. Technical solutions to solve this kind of problem are available (e.g. gazetters or thesauri) and those will be taken into account in a next step.

If there is a reason to doubt the semantics of a data object or if a researcher needs to investigate the semantics of the source data, a reference to the source database in its original model and data format is always provided for each data object of the integrated database.

4.1 Archaeological Domain

For each source database, the semantic mapping to the internal model is formulated in Python code. At first a *reader* for the file formats (mostly Excel and CSV) of the file based databases (CalPal, Stage3 and internal collections) was implemented. This reader maps the schema of the source dataset to a Python object, which is then mapped into RDF representation. The created RDF graph is then written into the central RDF store.

For the NESPOS database, a website *scraping* script was developed in order to collect the archaeological information from the public space of the NESPOS web portal. Until now, the collected data is site based. This may be extended to artefact records in the future. The RDF mapping is implemented in an additional step, the dataset also gets integrated by writing the resulting RDF graph into the central RDF store.

Each major entity (*Artefact*, *Site*, *SiteAttribution*) of the integrated database will be represented by a dynamically created webpage, accessible under its RDF URI within the CRC806-Database web portal. The webpage will contain all available information about the object, including references and links to the source database.

4.2 Palaeoenvironmental Domain

In the case of the palaeoenvironmental databases, the integration process is different from the approach applied to archaeological databases. The datasets of the archaeological domain are spatially mainly point (site) based, whereas datasets of the palaeoenvironmental data domain are mainly represented as discrete spatial fields (grids). This leads to a more GIS-based integration approach.

Similar to the integration process for the archaeological data, a custom reader for each dataset is implemented in a Python program. In contrast to the process for the archaeological domain, each dataset is transformed into a common GIS data format (GeoTiff or Shapefile) in addition to the RDF mapping.

In further contrast to the archaeological integration, not every data object of the source is mapped into RDF, but only important metadata and the references to the derived GIS datasets containing the values of all the data objects of the source. This is due to the nature of the palaeoenvironmental data, mostly represented in spatial grids, with n (see equation 1) values per grid node g, which results in $g \cdot n$ data objects per variable, easily resulting in a very large number of data objects to be mapped in RDF.

For each integrated palaeoenvironmental dataset, a website with all available information about the data and references to its source will be available from its RDF URI within the CRC806-Database web portal.

5 RESULTS

The main results of the presented work are i) a comprehensive integrated palaeoenvironmental and archaeological database, ii) a semantic web-enabled archaeological, and iii) palaeoenvironmental data model, as well as iv) shared spatial and temporal semantic models.

5.1 Archaeological Model

The archaeological model (see Figure 2) comprises three main objects: *Artefacts*, *Sites* and *SiteAttribution*. Most of the archaeological datasets integrated so far into the database and the underlying model are based on records relating to artefacts. *Artefacts* are located by a reference 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. For these kind of records, the object *SiteAttribution* was developed. It has the added ability of being able to 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 enables site based analysis.

Furthermore, references to the source databases, and in case of the NESPOS datasets, links to the webpages of the sites within the NESPOS website, are provided within the RDF graph model.

5.2 Palaeoenvironmental Model

As described above, not all data objects of the palaeoenvironmental source datasets are modeled in RDF. Thus, the model concentrates on the relevant parameters to provide integrated spatio-temporal filtering, and information about the present environmental variables in the datasets.

Each dataset (see Figure 3) is spatio-temporally referenced within the model, facilitating the shared temporal and the shared spatial

Figure 2: Generalized graph representation of the archaeological data model.

model. The palaeoenvironmental datasets are temporally defined by palaeoenvironmental periods, dates or events (OIS-Stages, glacial periods, Heinrich events etc.), which translate into time ranges or in case of events and dates into points in time.

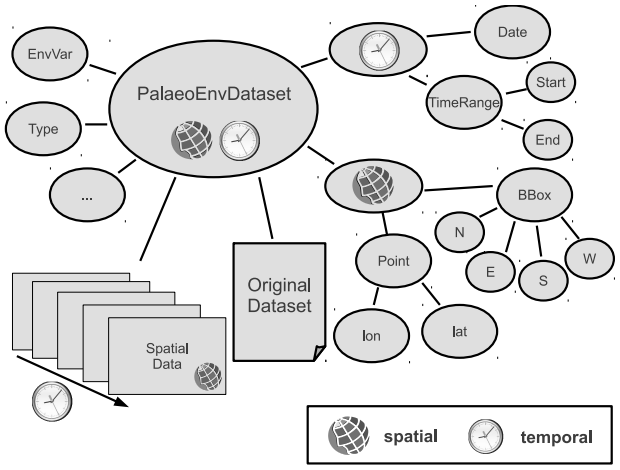

Figure 3: Generalized graph representation of the palaeoenvironmental data model and also simplified versions of the temporal and spatial models.

Spatially, the datasets are integrated with geographic coordinates (WGS84) boundingboxes of the derived GIS-datasets, described by the shared spatial model. Relevant meta information about the dataset, such as environmental variables and references to the original source datasets, are also represented in the RDF model.

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 sematics of the datamodel. The spatial data is stored internally in a processed GIS-data format, as well as in the original data format with a reference to the source database.

5.3 Shared models

The developed models share the same spatio-temporal model. The right side of Figure 3 shows an abstracted graph showing the simple spatial and temporal model. To describe the spatial and temporal extent of the entities, it is subdivided into a spatial and a temporal model. These models will be continually semantically extended and refined. In particular, a semantic integration with existing spatio-temporal ontologies is planned to strengthen the interoperability of the models. This semantic integration is implemented by formulating *OWL:sameAs* instances between individuals in the developed domain ontologies and existing vocabularies.

Spatial model: The spatial model is at this stage rather simple and basically represents spatial points, defined by a WGS84 Lat/Long coordinate. Spatial fields, which are defined by a simple bounding box, are represented by the bounding North, East, West and South WGS84 ordinates.

The WGS84 semantics of the spatial model are in alignment with the *Basic Geo (lat/long) Vocabulary* by the W3C Semantic Web Interest Group (2003), where the bounding box is defined by two coordinates, the minimum longitude and minimum latitude point and the maximum longitude and maximum latitude point.

Temporal model: The temporal model is slightly more complex than the spatial model, because it additionally deals with defined names of events and periods, which translate into periods and dates. Events translate into a point in time or date, and periods translating into time ranges, which are defined by a start and an end date.

Consequently, the model defines dates that are given in years BP (before present) as basic entities. Further investigation of the semantic integration of different time reference systems used in source datasets (e.g. BP, calBP, BC, etc.) will be undertaken and ontologically formulated within the model. At the moment, this is done manually during the formulation of the semantic mapping (see Section 4 and Figure 1). Because the representation of dates are simple integer values, the duration d of a time range is a simple substraction of start date sD minus end date eD ($d = |sD - eD|$).

6 INTEGRATED SPATIO-TEMPORAL VISUALIZATION AND ANALYSIS

The heterogeneous palaeoenvironmental and archaeological data is spatio-temporally integrated into a central RDF store and thus facilitates the implementation of integrative analysis and visualization applications.

The CRC806-Database architecture for accessing the integrated databasis is shown in Figure 4. The interfaces and applications access the integrated data through a SPARQL (Prud'hommeaux and Seaborne, 2008) endpoint, which operates as query engine on top of the central RDF store. The only access interface an application has to the integrated database is the SPARQL endpiont. This ensures that the application is independent from further federations of additional data sources to the RDF graph, even if additional semantics are added.

6.1 Interfaces

Thus, the integrated data can be accessed by several interfaces within the CRC806-Database web portal including i) a data catalog and search interface (web based SPARQL frontend), ii) a WebGIS and iii) an Exhibit timeline and facet-browsing application, iv) via direct access to the SPARQL endpoint and v) via OGC webservices (WFS, WMS, WCS).

All GIS datasets of the integrated databasis are accessible from the *CRC806-Database SDI* via OGC webservice (OWS) interfaces, implemented using Open-Source-Software Mapserver (`http://www.mapserver.org`) and MapProxy (`http://www.mapproxy.`

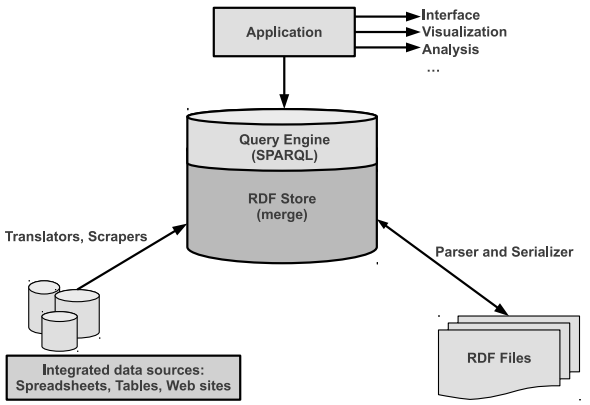

Figure 4: Architecture for accessing the integrated databasis (modified from Allemang and Hendler (2011)).

org) technology. The geospatial datasets are also directly downloadable in Shapefile or GeoTiff format from the *CRC806-Database web portal* data directory via HTTP file download.

The *CRC806-Database web portal* is a Typo3 CMS-based web application and hosted at the server infrastructure of the regional computing centre of the University of Cologne (RRZK). The data catalog and search interface for the integrated data is implemented in a custom Typo3 Extension, which allows the implementation of interfaces to construct SPARQL requests and display the results from within Typo3. Through this interface it is also possible to export results of SPARQL queries as RDF/XML serializations and in CSV and Excel format.

6.2 Example Queries

The integrated RDF model allows to filter the integrated data base with queries, which was not possible before the integration of the data. Some example queries on the integrated database are:

- Select all datasets located in South Spain from the LGM time intervall.
- In which time intervalls and in which spatial areas are artefacts from the solutrean culture present in the database?

It enables to filter all datasets of the integrated databasis spatially and temporally in addition to thematical filters. This was not possible before in a consistent integrated way.

6.3 Visualization

For the interactive visualization in a spatial context, a WebGIS is provided (see Figure 5) accessing the OGC interfaces of the *CRC806-Database SDI*. The WebGIS is implemented using the open source JavaScript framework GeoExt (`http://www.geoext.org/`).

A further interactive interface for temporal (timeline) visualization and structured faceted browsing of the integrated data is provided using the open source Exhibit JavaScript framework (Huynh et al., 2007). Faceted browsing helps exploring the data based on more than one axis (e.g. a search term) by applying multiple filters in faceted classification system (Huynh et al., 2007).

The visualization of the integrated data in desktop/client applications is possible through accsess of the OWS interfaces or the file

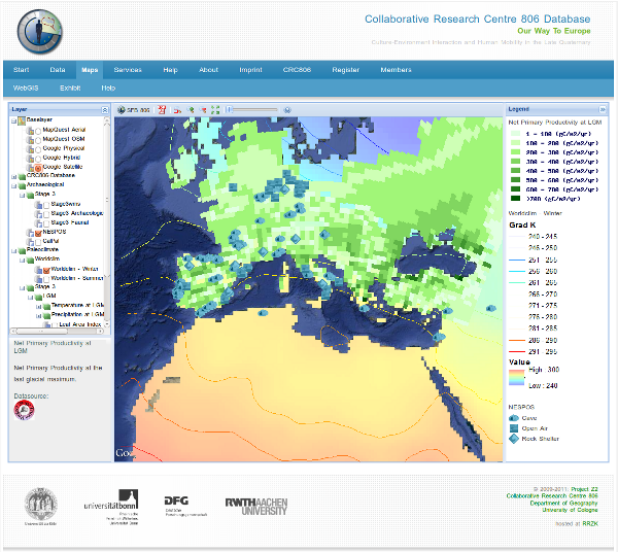

Figure 5: Screenshot of the WebGIS interface, showing a Google Maps *Aerial* baselayer overlayed with the *mean LGM winter temperatures* from PMIP II, and overlayed with the *LGM net primary productivity* from Stage3 and the NESPOS Sites.

based datasets, which can be accessed from the web portal. Thus, many possibilities of custom analysis and visualizations are additionally enabled by the web based interfaces of the described system.

7 CONCLUSION AND OUTLOOK

Originally, a top-down approach was considered, which would have integrated the given data into an upfront-developed model. This approach was abandoned due to its limited flexibility and its susceptibility to error. This led to the adoption of a bottom-up approach, which builds the model from the semantics of the integrated data sources. This approach has several advantages, such as flexibility and extendibility. 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 can be semantically mapped to existing models, which provide a *semantic overlap* (Allemang and Hendler, 2011), by the definition of *OWL:sameAs* statements. This enables the definition of mappings to representations in existing models such as CIDOC-CRM (Doerr, 2003) and thus declares data in those models, with a mapping to the CRC806-Databases model, interoperable with the CRC806-Database model. This semantic referencing and linkage will be applied to as many available external models and ontologies as possible, to strengthen the semantic interoperability of the presented archaeological and palaeoenvironmental models.

The main result of this work is the integrated data basis and the data models derived from the integration process. This data basis will help to answer and inspire a wide range of research questions within the CRC806, and possibly within the research community as a whole. Further datasets will be integrated into the CRC806-Database in the future. Users of the CRC806-Database will be able to suggest further datasets for integration.

The interfaces and applications provided by the CRC806-Database so far are just a small fraction of possible applications, that can be implemented on top of the data basis. Further applications are

for example environmental and archaeological predictive models such as archaeological *site catchment* and *site prediction* analyses or for example *palaeo climate classifications*. It is already planned to provide OGC Web Processing Service (WPS) interfaces for implementation of such models, to integrate those for interactive exploration within the CRC806-Databse WebGIS application.

The CRC806-Database web portal and SDI (`http://crc806db.uni-koeln.de`) will be launched in summer 2012. This first version of the web portal will implement the data management aspect, including a secure research data archive and some first web based interfaces and applications to the integrated data basis.

ACKNOWLEDGEMENTS

The authors would like to thank the creators of the integrated datasets for the publication of their research results and thus providing them to the community and the research project presented in this contribution. Furthermore we thank the German Research Foundation (DFG) for funding the CRC806.

References

Allemang, D. and Hendler, J., 2011. Semantic web for the working ontologist: modeling in RDF, RDFS and OWL. 2nd edn, Morgan Kaufmann Publishers/Elsevier.

Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F. and Stein, L. A., 2004. OWL Web Ontology Language Reference. Technical report, W3C.

Braconnot, P., Otto-Bliesner, B., Harrison, S., Joussaume, S., Peterschmitt, J.-Y., Abe-Ouchi, A., Crucifix, M., Driesschaert, E., Fichefet, T., Hewitt, C. D., Kageyama, M., Kitoh, A., Lan, A., Loutre, M.-F., Marti, O., Merkel, U., Ramstein, G., Valdes, P., Weber, S. L., Yu, Y. and Zhao, Y., 2007. 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), pp. 261–277.

Bradtmöller, M., Pastoors, A., Slizewski, A. and Weniger, G.-C., 2010. NESPOS- A digital archive and platform for Pleistocene archaeology. In: C. Curdt and G. Bareth (eds), Proceedings of the data management workshop, Kölner Geographische Arbeiten, Vol. 90, University of Cologne, pp. 13–18.

Curdt, C., Hoffmeister, D., Jekel, C., Brocks, S., Waldhoff, G. and Bareth, G., 2011. TR32DB Management and visualization of heterogeneous scientific data. In: Proceedings of the 19th International Conference on Geoinformatics, Shanghai, China.

DFG, 1998. Proposals for Safeguarding Good Scientific Practice - Recommendations of the Commission on Professional Self Regulation in Science. Technical report, Deutsche Forschungsgemeinschaft, Weinheim, Germany.

Doerr, M., 2003. The CIDOC-CRM - An ontological approach to semantic interoperability of metadata. AI Magazine 24, pp. 2003.

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

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

Hoelzmann, P., Jolly, D., Harrison, S., Laarif, F., Bonnefille, R. and Pachur, H.-J., 1998. 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), pp. 35–52.

Huynh, D. F., Karger, D. R. and Miller, R. C., 2007. Exhibit: lightweight structured data publishing. In: Proceedings of the 16th international conference on World Wide Web, WWW '07, ACM, New York, NY, USA, pp. 737–746.

Isaksen, L., Martinez, K., Gibbins, N., Earl, G. and Keay, S., 2009. Linking archaeological data. In: Computer Applications and Quantitative Methods in Archaeology conference, CAA.

Kauppinen, T., Mantegari, G., Paakkarinen, P., Kuittinen, H., Hyvnen, E. and Bandini, S., 2010. Determining relevance of imprecise temporal intervals for cultural heritage information retrieval. International Journal of Human-Computer Studies 68(9), pp. 549 – 560.

Kavouras, M. and Kokla, M., 2008. Theories of Geographic Concepts : Ontological Approaches to Semantic Integration. CRC Press, Taylor & Francis, Boca Raton, FL, USA.

Klyne, G. and Carroll, J. J., 2004. Resource Description Framework (RDF): Concepts and Abstract Syntax. Technical report, W3C. Online: http://www.w3.org/TR/rdf-concepts/.

Martinez, K. and Isaksen, L., 2010. The semantic web approach to increasing access to cultural heritage. In: C. Bailey and H. Gardiner (eds), Revisualizing Visual Culture, Ashgate, pp. 29–44.

Prud'hommeaux, E. and Seaborne, A., 2008. SPARQL Query Language for RDF. Technical report, W3C. Online: http://www.w3.org/TR/rdf-sparql-query/.

Segaran, T., Evans, C. and Taylor, J., 2009. Programming the Semantic Web. O'Reilly, Sebastopol, CA, USA.

Sellis, T., Koubarakis, M., Frank, A., Grumbach, S., Gting, R. H., Jensen, C., Lorentzos, N., Manolopoulos, Y., Nardelli, E., Pernici, B., Theodoulidis, B., Nectaria, T. N., Schek, H.-J., and Scholl, M., 2003. Spatio-temporal databases: the CHOROCHRONOS approach. Lecture Notes in Computer Science 2520, Springer, Berlin, New York.

van Andel, T. and Davies, W., 2003. 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.

Visser, U., 2004. Intelligent information integration for the Semantic Web. Lecture Notes in Artificial Intelligence, Vol. 3159, Springer-Verlag, Berlin, Heidelberg.

W3C Semantic Web Interest Group, 2003. Basic Geo (WGS84 lat/long) Vocabulary. Online: http://www.w3.org/2003/01/geo/.

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

Willmes, C. and Bareth, G., 2012. A data integration concept for an interdisciplinary research database. In: A. Degbelo, J. Brink, C. Stasch, M. Chipofya, T. Gerkensmeyer, M. I. Humayun, J. Wang, K. Brolemann, D. Wnag, M. Eppe and J. H. Lee (eds), Proceedings of the Young Researchers forum on Geographic Information Science - GI Zeitgeist, ifgiPrints 44, Akademische Verlagsgesellschaft AKA, Heidelberg, pp. 67 – 72.