Jahresversammlung SGA/SSA und AGHAS
16 November 2019

RDFBones: A framework for the standardisation of osteological data

Felix Engel, Stefan Schlager
University of Freiburg
Biological Anthropology – Faculty of Medicine

Text version of the podium presentation.

Contents

Introduction

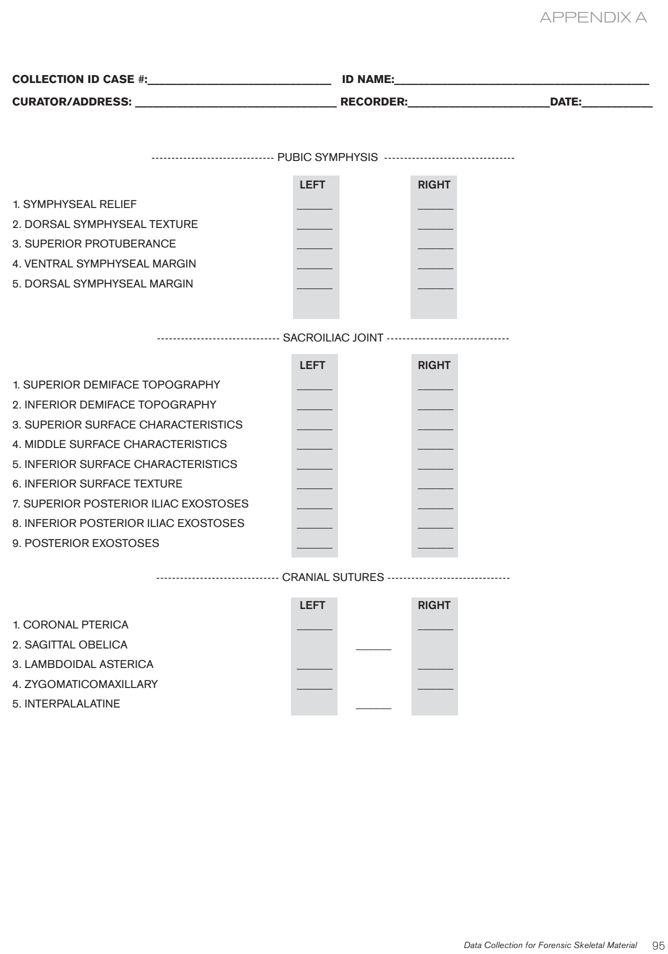
This paper relates results from a three-year project funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) which developed a digital standard for scientific information produced from osteological research collections. It is the basis for a follow-on project that will start in December 2019.

The paper will set off by summarising approaches to data standardisation during the last couple of years and then will take a brief look at semantic data modelling, a practice adopted in other life sciences. Finally, the digital meta-standard RDFBones will be introduced which applies semantic data modelling to osteological research in biological anthropology. There will also be a short outlook on the implementation of RDFBones as planned in the research project ahead.

Data Standardisation in Osteology

The initial requirement for research data standardisation is standardised methodology. As a consequence, standards include instructions on data production, e. g. how analyses are

This research was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) under the reference WI 863/9-1.



APPENDIX A

COLLECTION ID CASE #:________________________ ID NAME:________________________________

CURATOR/ADDRESS: ________________________ RECORDER:__________________DATE:__________

PUBIC SYMPHYSIS

	LEFT	RIGHT
1. SYMPHYSEAL RELIEF	_____	_____
2. DORSAL SYMPHYSEAL TEXTURE	_____	_____
3. SUPERIOR PROTUBERANCE	_____	_____
4. VENTRAL SYMPHYSEAL MARGIN	_____	_____
5. DORSAL SYMPHYSEAL MARGIN	_____	_____

SACROILIAC JOINT

	LEFT	RIGHT
1. SUPERIOR DEMIFACE TOPOGRAPHY	_____	_____
2. INFERIOR DEMIFACE TOPOGRAPHY	_____	_____
3. SUPERIOR SURFACE CHARACTERISTICS	_____	_____
4. MIDDLE SURFACE CHARACTERISTICS	_____	_____
5. INFERIOR SURFACE CHARACTERISTICS	_____	_____
6. INFERIOR SURFACE TEXTURE	_____	_____
7. SUPERIOR POSTERIOR ILIAC EXOSTOSES	_____	_____
8. INFERIOR POSTERIOR ILIAC EXOSTOSES	_____	_____
9. POSTERIOR EXOSTOSES	_____	_____

CRANIAL SUTURES

	LEFT		RIGHT
1. CORONAL PTERICA	_____		_____
2. SAGITTAL OBELICA		_____	
3. LAMBDOIDAL ASTERICA	_____		_____
4. ZYGOMATICOMAXILLARY	_____		_____
5. INTERPALALATINE		_____	

Data Collection for Forensic Skeletal Material 95

Figure 1. A page from the osteological data standard 'Data Collection Procedures for Forensic Skeletal Material' (Langley, Jantz, Ousley, Jantz and Milner 2016, 95).

to be conducted. Ideally, these instructions are backed up by scientific publications. But standardised methodology alone does not constitute a data standard as it does not describe how results from osteological investigations are to be recorded.

We understand a data standard to consist of a series of well-defined data items describing results from reproducible observations, calculations based on these results and conclusions drawn from other data items. A formal data standard needs to be openly accessible, e. g. as a scientific publication. Such specifications define data models that can be implemented in databases to the effect that independently produced information can be easily merged into a single dataset. Data standards are typically implemented in data recording sheets as shown in figure 1.

Figure 2 shows a timeline of major osteological data standards in biological anthropology (Brickley and McKinley 2004; Buikstra and Ubelaker 1994; Harbeck 2019; Langley et al. 2016; Moore-Jansen, Ousley and Jantz 1994; Steckel, Larsen, Sciulli and Walker 2018;

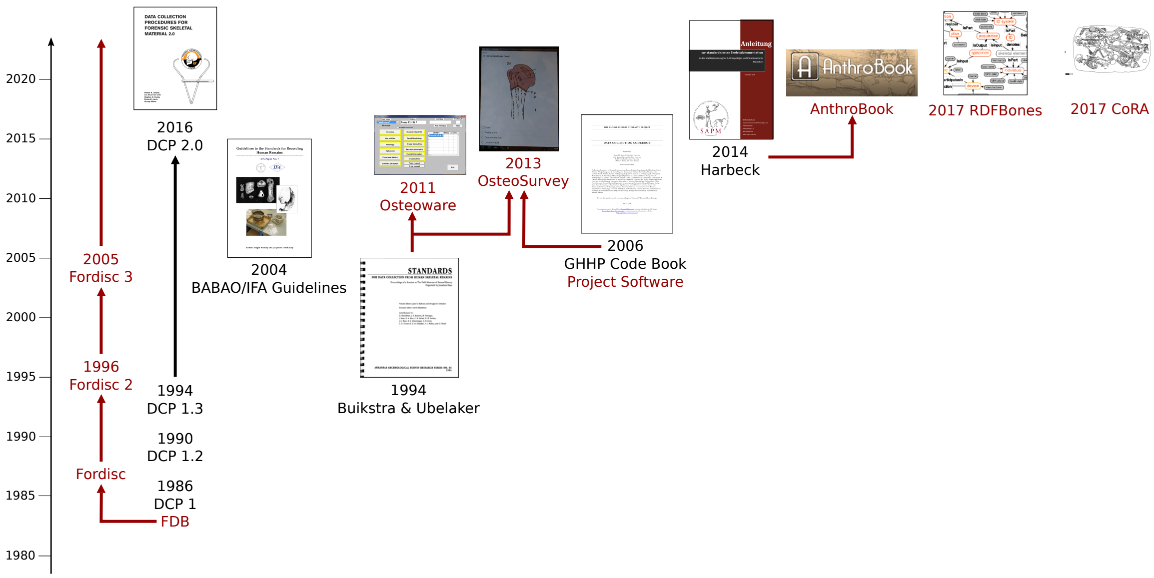


Figure 2. Timeline of osteological data standards and related software implementations (marked in red).

Trautmann 2019; Trautmann and Harbeck 2019). The number of available standards has increased since the mid-1980ies and most of them have been implemented in software tools supporting data acquisition and analysis (Austin 2017; Dudar et al. 2017; Engel and Schlager 2019; Jantz and Ousley 2005; Kaltenthaler, Lohrer, Kröger and Obermaier 2017; Lynch and Stephan 2018). The 'Data Collection Procedures for Forensic Skeletal Material' (Langley et al. 2016; Moore-Jansen et al. 1994) stand out as the only one to have appeared in a series of updated versions.

These standardisation efforts follow a number of different motivations. They ensure and enforce quality standards in the analysis of osteological material, e. g. with the documentation of material that is going to be restituted or reburied and, therefore, only available for a limited time. Another context of quality enforcement is contract work performed by freelance researchers. As noted above, data standards also support the pooling of independently created dataset into larger databases. This is a basis for collaborative projects involving several researchers. Finally, the development of scientific software requires sound data models which also arise from data standardisation.

An additional objective of data standardisation might be the formal declaration of research designs. But this would require much more complex data models and has not been a consideration so far.

There is a considerable overlap between the existing data standards, i. e. data items have been recreated in an attempt to deliver improvement to certain areas of research. As a consequence, different standards contain compatible information without supporting data pooling due to diverging data models. A possible solution would be the formulation of an overarching standard serving all use cases. Why this approach is problematic is illustrated by the cartoon 'Standards' by Randall Munroe (figure 3).

It is striking that after more than 30 years of data standardisation, the production

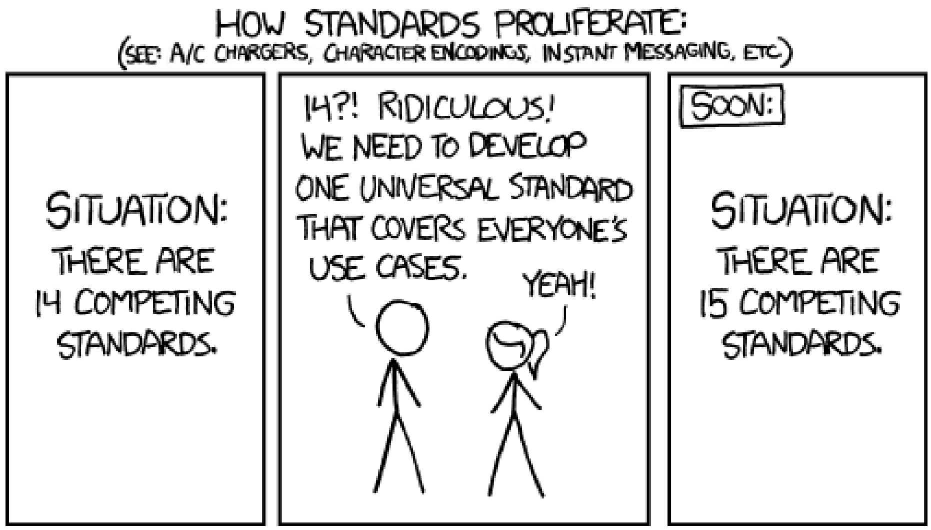


Figure 3. xkcd 927, 'Standards' by Randall Munroe (https://xkcd.com/927/).

of standardised data sets ready for reuse in other research contexts has not become a standard procedure in biological anthropology. Instead of constructing better and more comprehensive data standards it might be more rewarding to consider possible reasons why existing standards are not being used.

The formulation of existing data standards mostly followed the logic of relational databases that required a database to be established before it could be filled with information. Later changes to the data model came at a considerable cost making database systems unnecessarily complex and difficult to query. This authoritative approach comes with a number of problems. As data models are fixed, even minor modifications or extensions to incorporate recently developed methods or new research questions and designs are hardly possible. As a consequence, data standards – or at least the databases implementing them – become outdated because they cannot consider scientific progress. Also, it is quite unclear which authority should define data standards and, thereby, tell other researchers how to conduct investigations.

Another difficulty is the interdisciplinary nature of biological anthropology. Many research projects are not refined to osteological investigations alone but analyse their results on the background of some external information. Traditionally, this requires pulling data from several databases and setting up a new project-related database. Apart from being rather work intensive, this approach has the disadvantage that the obtained information is separated from its original database to the effect that corrections and additions to the original dataset are not reflected. This leads to research data becoming quickly outdated and limits their re-usability.

To conclude, another authoritative data standard, no matter how concise, would not bring a major improvement. Instead, it is necessary to acknowledge and account for the diversity of research data in biological anthropology. For inspiration on how to achieve this it might be rewarding to take a look at how data standardisation is handled in other disciplines.

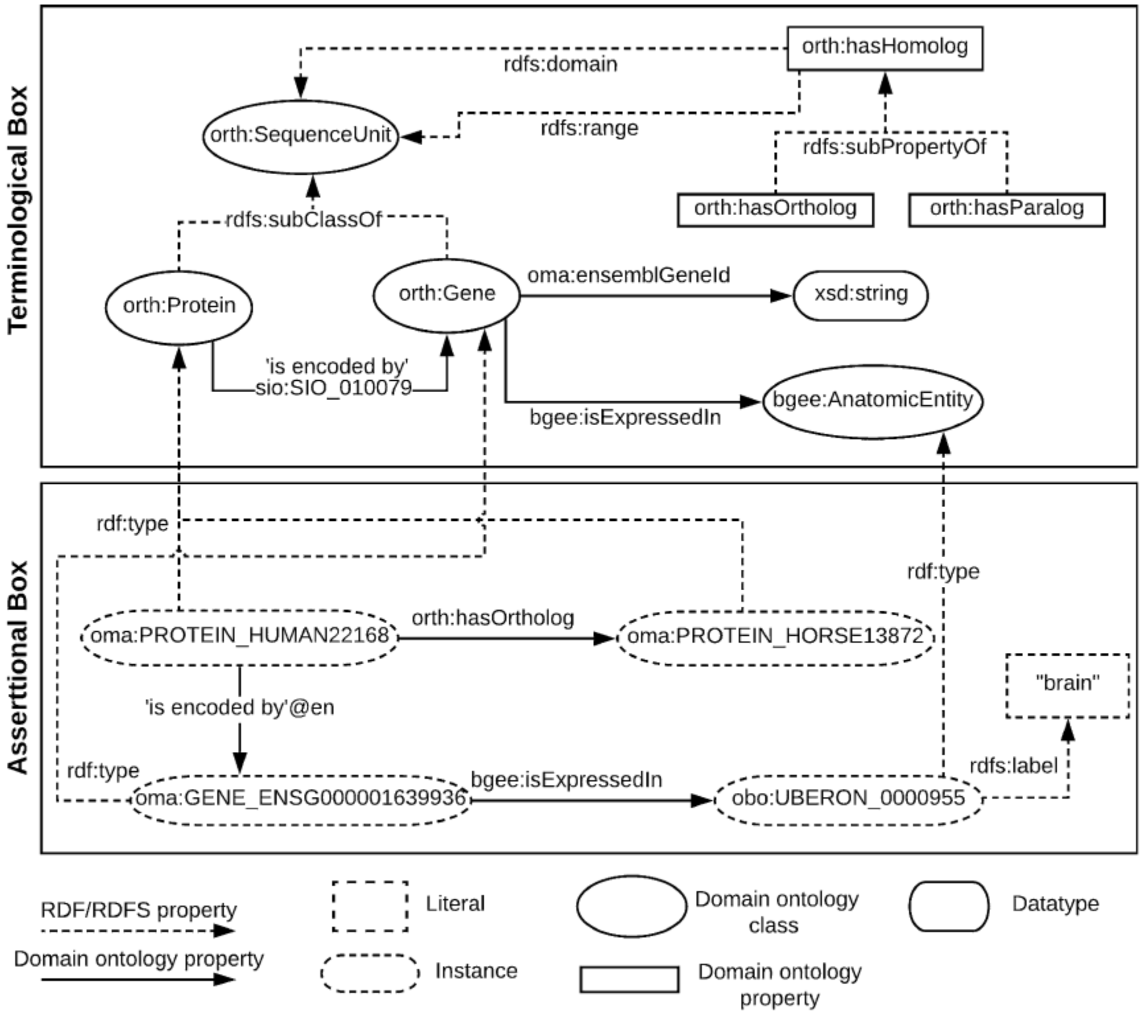


Figure 4. An example of semantic modelling of genetic information (Sima 2019, fig. 10).

Semantic Data Modelling in the Life Sciences

An approach widely used in many life sciences is semantic modelling of research data (figure 4). An accessible introduction to the concept is provided by Sima (2019). Here it suffices to note that the data model consists of precisely defined data items represented by bubbles and boxes that are linked by a number of specific types of relations represented as arrows. The upper panel defines how the encoding of proteins by genes and their expression in organisms works in general. This body of general knowledge about genetics is referred to as an ontology. The lower panel fills this general model with more specific information, i. e. the encoding of a specific protein in a specific gene expressed in the brain, a specific anatomical entity.

Ontologies covering various domains of knowledge are constantly curated by scientists and provided through dedicated online platforms. Figure 5 shows the OBO Foundry website as an example (Smith et al. 2007).

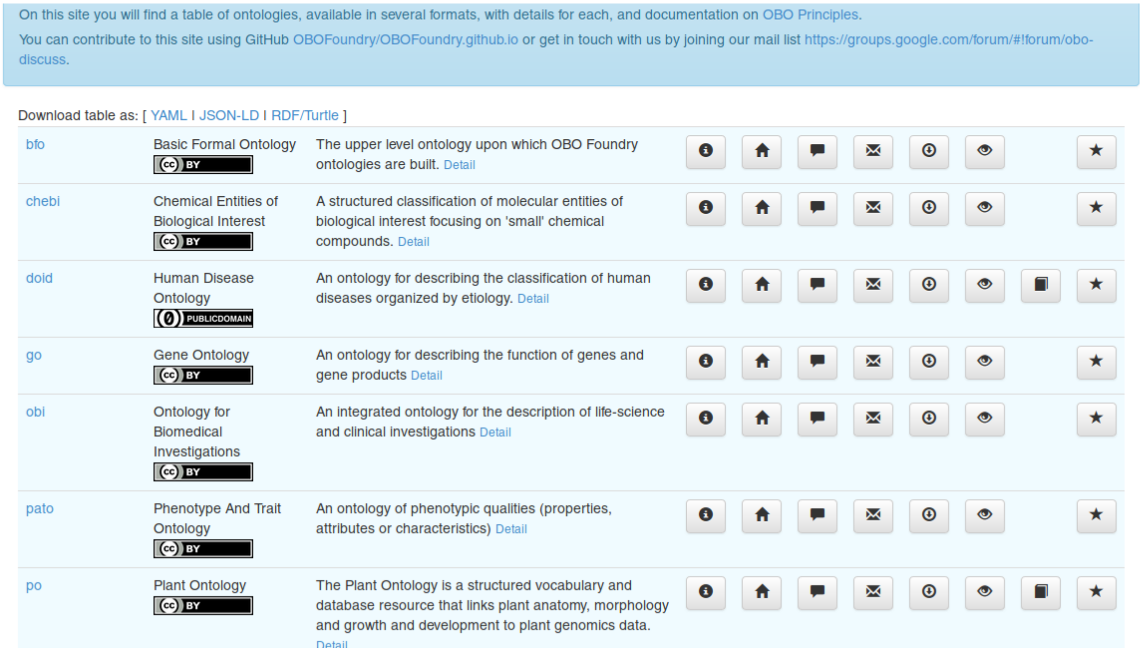


Figure 5. Biolomedical ontologies provided on the OBO Foundry website (http://obofoundry.org).

How a particular research project might use semantic data modelling is illustrated in figure 6. Several existing databases (marked in green) are mapped onto a number of reference ontologies (marked in red, pink and violet) by relating their data fields to equivalent ontology items. Following these connection lines, queries based on the ontologies draw data from the disparate data sources. These queries can be executed over and over again, even while the feeding databases are continuously updated and expanded.

RDFBones and Application in Biological Anthropology

RDFBones is an ontology applying the concept of semantic research data modelling to osteology. Its purpose is to link a number of domain ontologies covering various aspects of osteological research in order to form a general model of this type of research. (figure 7).

The general model of osteological research is referred to as the core ontology in RDFBones. It can be concretised through ontology extensions which define data models implied by individual research designs while using elements from the core ontology (figure 8). In this way, research projects can structure their data according to their specific needs and also include data items that are particular to their research questions, e. g. contextual data. At the same time, they contribute data encoded by data items from the core ontology to the overall pool of osteological information.

While this concept might present a viable solution to the problems with data standardisation in biological anthropology identified above, the majority of anthropological researchers will be at a loss how to put semantic data modelling into practice in their everyday work. Obviously, researchers in other disciplines produce and interact with semantic

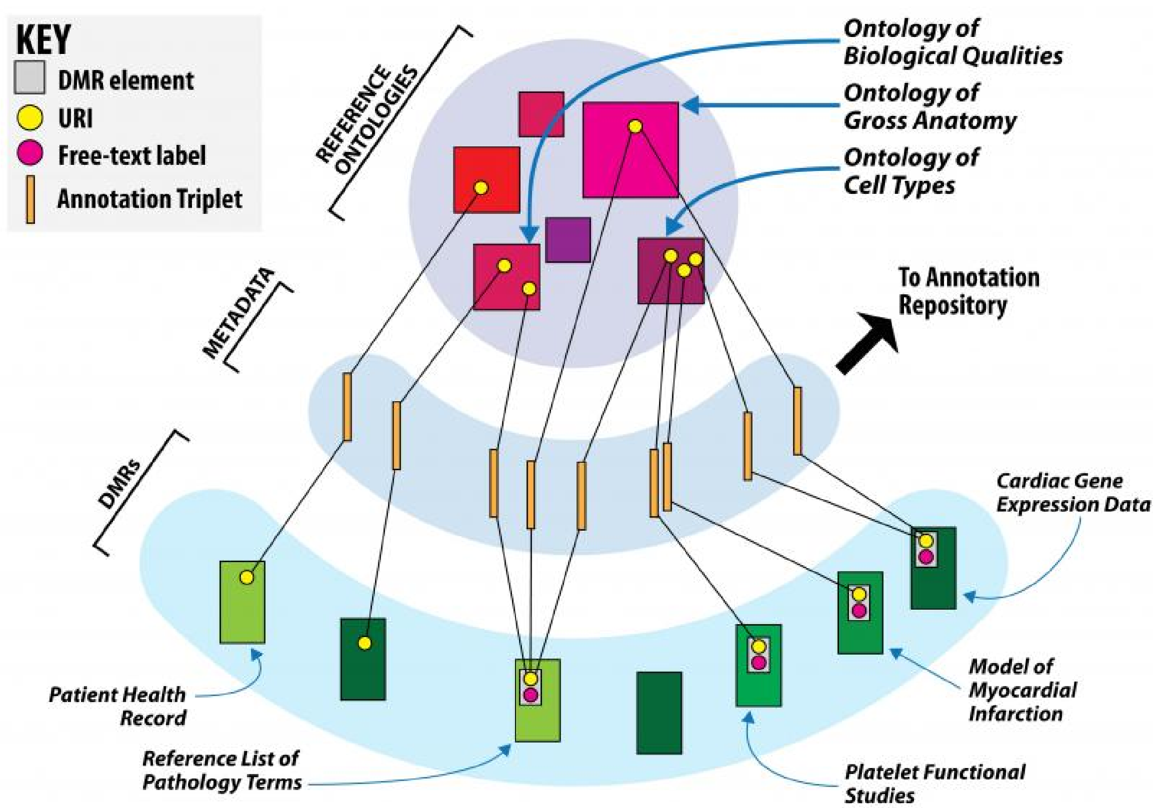


Figure 6. Example of semantic data modelling employed to query data across several databases (project 'Ricordo: Interoperable Anatomy and Physiology', http://www.ricordo.eu).

research data in many ways that cannot be presented here.

One attempt to make semantic modelling of osteological research data a viable option for biological anthropologists will be made by an upcoming research project at Freiburg University. Its objective is to develop an online information system with the working title AnthroGraph which will provide researchers with data input forms, just as the ones provided by established software tools for data acquisition. With AnthroGraph however, researchers will be able to configure the system according to their individual needs by loading RDFBones extensions. Usage will be similar to other database applications but AnthroGraph will create semantic data models from the input that can be pooled with information entered through other ontology extensions. AnthroGraph will be built by adapting the software framework ResearchSpace which is developed by the British Museum (figure 9). While the original ResearchSpace system models research data related to objects from the British Museum collections, AnthroGraph will be specifically configured for osteological data.

Semantic research data modelling has a number of advantages and may solve some of the problems with existing approaches to the standardisation of osteological data in biological anthropology. First of all, it can be used to pool data from existing bodies of structured information without having to recode the information. With large quantities of

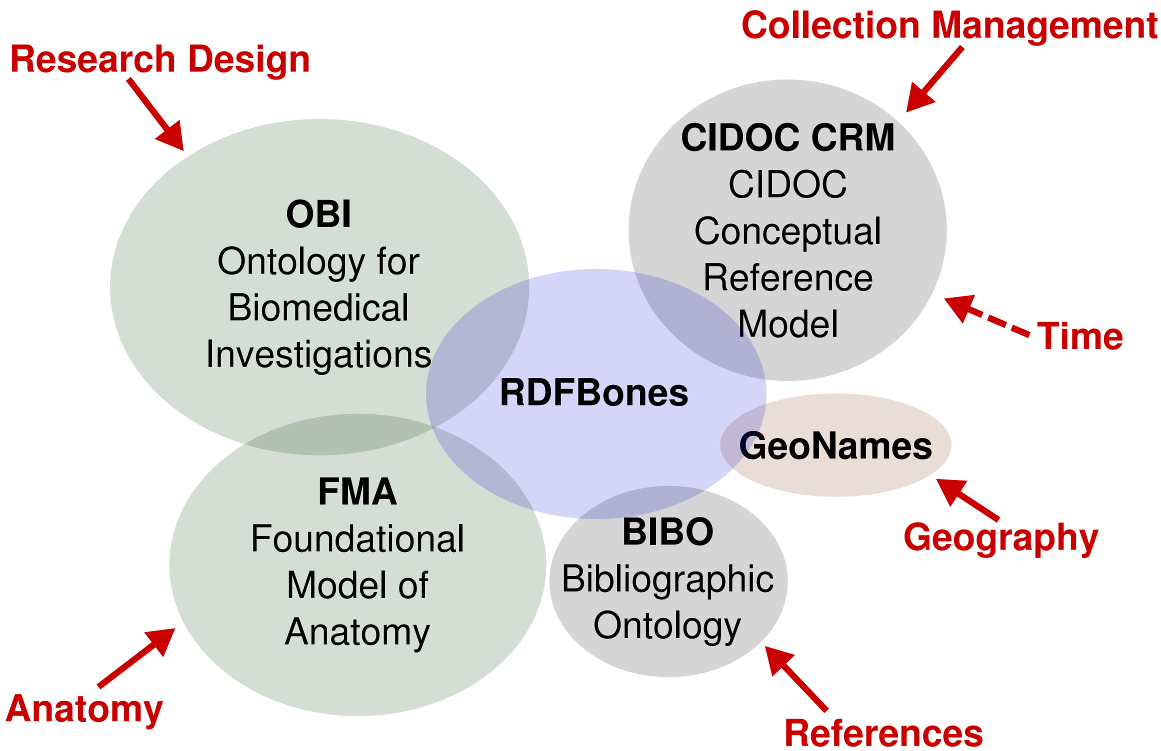


Figure 7. RDFBones and contributing domain ontologies covering various aspects of osteological research.

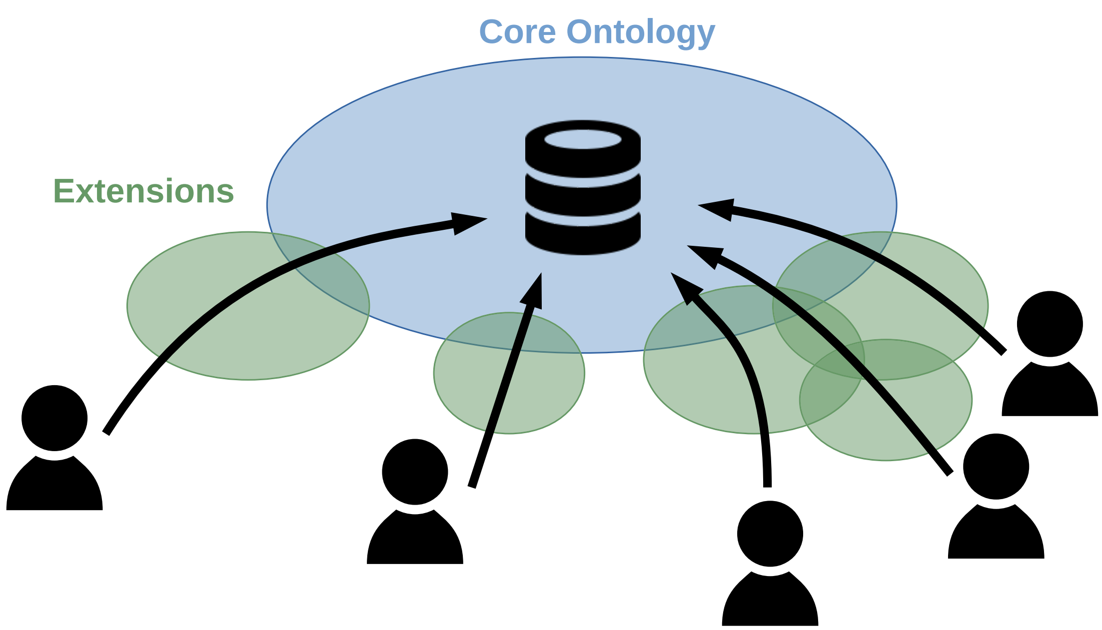


Figure 8. Modular structure of RDFBones with core ontology and extensions.

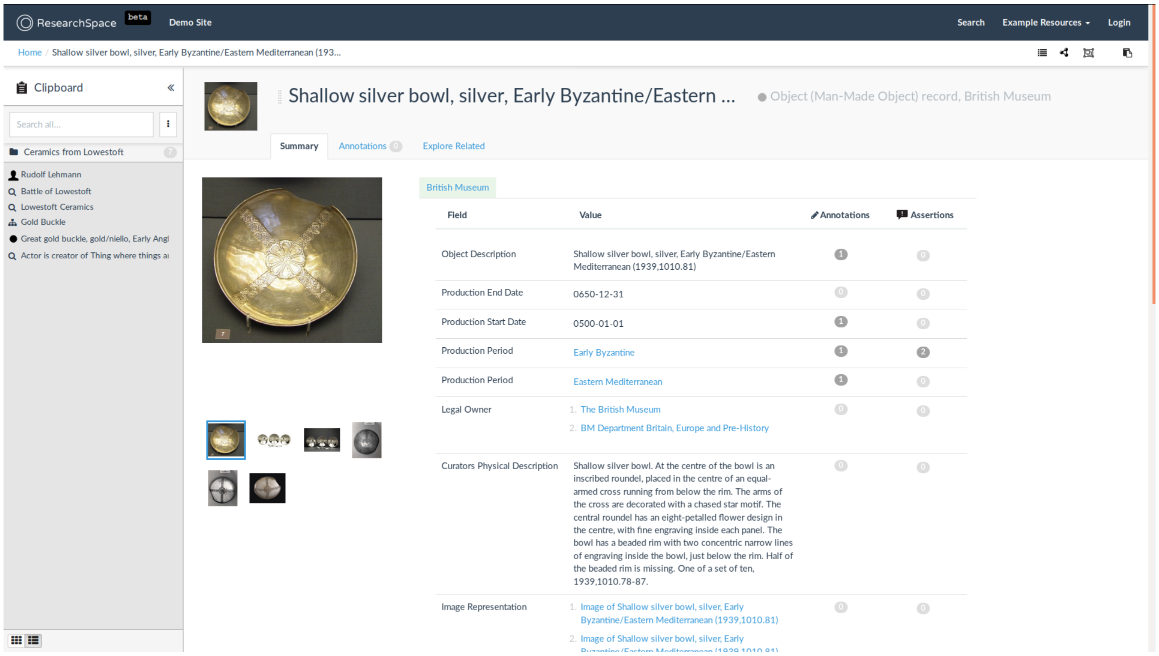


Figure 9. Screenshot of the ResearchSpace software (https://www.researchspace.org).

data this saves a lot of work effort but also avoids transformation errors and the creation of redundant information. Specific research designs as well as new research topics and methods can be incorporated by creating dedicated ontology extensions. As these are all based on the RDFBones core ontology their data models retain a certain degree of compatibility and there are precise definitions of which data items are compatible and which are not. As the creation of new ontologies does not compromise the overall aggregation of compatible data, standards do not need to be enforced but can organically evolve.

But semantic data modelling also offers improvements for the analysis of osteological data. As RDFBones uses elements from established ontologies, data can be analysed on the background of these reference ontologies, e. g. taking into account entire models of human anatomy. Extensions may also link to ontologies that are not represented in the RDFBones core ontology, opening up additional knowledge contexts. This quality of semantic data modelling offers support for any kind of interdisciplinary research.

On the other hand the introduction of semantic data modelling in biological anthropology has to face challenges rooted in the condition of the discipline and its institutions. It would require an increased interest in the provision of standardised research data than currently present. The relatively small scientific community of biological anthropologists will find it harder to foster work force for the curation of data standards and research databases than it is the case in other life sciences. Therefore, it might be necessary to recur to existing infrastructures in other disciplines or with institutions dedicated to research data management in general.

References

Austin, A. E. (2017). OsteoSurvey: An open-source data collection tool for studying commingled human remains. In *American journal of physical anthropology* (Vol. 162, p. 105). Hoboken: Wiley. doi: 10.1002/ajpa.23210

Brickley, M. & McKinley, J. I. (Eds.). (2004). *Guidelines to the standards for recording human remains* (No. 7). Institute for Field Archaeologists. Retrieved from `http://www.babao.org.uk/HumanremainsFINAL.pdf`

Buikstra, J. E. & Ubelaker, D. H. (Eds.). (1994). *Standards for data collection from human skeletal remains* (Vol. 44). Fayetteville: Arkansas Archaeological Survey.

Dudar, C., Ousley, S., Jones, E., Wilczak, C., Hefner, J., Gwyn, M. & Mulhern, D. (2017). Osteoware: Standardized skeletal documentation software at the smithsonian institution. In *American journal of physical anthropology* (Vol. 162, p. 167). Hoboken: Wiley. doi: 10.1002/ajpa.23210

Engel, F. & Schlager, S. (2019, September). RDFBones – making research explicit: an extensible digital standard for research data. *Anthropologischer Anzeiger*, *76*(3), 245–257. doi: 10.1127/anthranz/2019/0882

Harbeck, M. (2019, October). Anleitung zur standardisierten Skelettdokumentation in der Staatssammlung für Anthropologie und Paläoanthropologie München: [Computer software manual]. München. Retrieved from `http://www.sapm.mwn.de/anthropologie/images/pdfs/Richtlinien/SAPM_Richtlinien_I_Erwachsene-komprimiert.pdf`

Jantz, R. & Ousley, S. (2005). Fordisc 3: Computerized forensic discriminant functions. version 3 [Computer software manual]. Knoxville.

Kaltenthaler, D., Lohrer, J.-Y., Kröger, P. & Obermaier, H. (2017). A framework for supporting the workflow for archaeo-related sciences: Managing, synchronizing and analyzing data. In B. Mitschang et al. (Eds.), *Datenbanksysteme fÃŒr business, technologie und web (btw 2017) - workshopband* (p. 89-98). Bonn: Gesellschaft fÃŒr Informatik e.V.

Langley, N. R., Jantz, L. M., Ousley, S. D., Jantz, R. L. & Milner, G. (2016). *Data collection procedures for forensic skeletal material 2.0.* Knoxville: Department of Anthropology, University of Tennessee.

Lynch, J. J. & Stephan, C. N. (2018, September). Computational tools in forensic anthropology: The value of open-source licensing as a standard. *Forensic Anthropology*, *1*(4), 228–243. doi: 10.5744/fa.2018.0025

Moore-Jansen, P. M., Ousley, S. D. & Jantz, R. L. (1994). *Data collection procedures for forensic skeletal material.* Knoxville: Department of Anthropology, University of Tennessee.

Sima, A. C. (2019). Semantic integration and enrichment of heterogeneous biological databases. In M. Anisimova (Ed.), *Evolutionary genomics: Statistical and computational methods* (Vol. 1910, pp. 655–690). New York: Humana. doi: 10.1007/978-1-4939-9074-0_22

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., ... Lewis, S. (2007, November). The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, *25*(11), 1251–1255. doi: 10.1038/nbt1346

Steckel, R. H., Larsen, S. C., Sciulli, P. W. & Walker, P. (2018, October). Data collection codebook. In R. H. Steckel, S. C. Larsen, C. A. Roberts & J. Baten (Eds.), *The backbone of Europe: Health, diet, work and violence over two millennia* (pp. 397–427). Cambridge: Cambridge University Press. doi: 10.1017/9781108379830

Trautmann, B. (2019, August). Anleitung zur standardisierten Skelettdokumentation in der Staatssammlung für Anthropologie und Paläoanthropologie München. Teil II: Leichenbrand (Second ed.) [Computer software manual]. München. Retrieved from http://www.sapm.mwn.de/anthropologie/attachments/article/249/RichtlinienSAPMLeichenbrand.pdf

Trautmann, B. & Harbeck, M. (2019, October). Anleitung zur standardisierten Skelettdokumentation in der Staatssammlung für Anthropologie und Paläoanthropologie München. Teil II: Leichenbrand. Teil III: Körpergräber von Kindern und Jugendlichen [Computer software manual]. München. Retrieved from http://www.sapm.mwn.de/anthropologie/attachments/article/249/RichtlinienSAPMLeichenbrand.pdf