Project Strategy

Aims and Opportunities for Action to Introduce Semantic Research Data Modelling in Biological Anthropology

Felix Engel, Stefan Schlager

11th September 2018

Felix.Engel@anthropologie.uni-freiburg.de

Contents

Nomenclature

AAA	American Anthropological Association
AAFS	American Academy of Forensic Sciences
AAPA	American Assosciation of Physical Anthropologists
DFG	Deutsche Forschungsgemeinschaft (German Research Foundation)
GfA	Gesellschaft für Anthropologie (Society for Anthropology)
GHHP	Global History of Health Project
HSC	Human Skeletal Collections (research project of the Biological Anthropology department of Freiburg University)
NFDI	Nationale Forschungsdateninfrastruktur (National Research Data Infrastructure)
RDF	Resource Description Framework
RDFBones	Digital standard for osteological research data
RfII	Rat für Informationsinfrastrukturen (Council for Scientific Information Infrastructures)
SAPM	Staatssammlung für Anthropologie und Paläoanaotomie München (State Collection for Anthropology and Paleoanatomy Munich)

1 Scope of this Document

From 2014 to 2017, researchers at the Biological Anthropology department of Freiburg University developed RDFBones, a digital data standard for research data emanating from osteological investigations in biological anthropology. The work was conducted within the ”Human Skeletal Collections (HSC)” project, funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG). During the project, a survey was conducted and several workshops and symposia held, yielding information on problems and requirements of research data management in biological anthropology, both in Germany and on an international level. It was also found that the RDFBones approach holds a potential to overcome current problems, address new challenges and bring about improvement to research data management in biological anthropology in the future.

With the intention to realise the potential of RDFBones and to prevent premature abandonment of this approach, the RDFBones work group submits a new proposal, ”Establishing Semantic Research Data Modelling in Biological Anthropology”, to the DFG, which is placed within the funding scheme ”e-Research Technologies”. This application proposes immediate steps for rendering RDFBones usable in research and demonstrating its capabilities to biological anthropologists.

This document presents a long-term perspective for the introduction of semantic research data modelling in biological anthropology, providing the larger framework within the proposal is formulated. It needs to be understood in the context of the proposal and knowledge of its project description is assumed.

2 Project Aims

Based on the argumentation outlined in section 1 of the project description, we propose to actively promote and establish semantic research data modelling in biological anthropology by employing the resource description framework (RDF). This will only be possible on the background of a growing awareness of research data management as a central element of research and the development of regular routines in the production and maintenance of research data. While these will improve data quality and reliability, they will also extend the workload of researchers and require additional infrastructures. Therefore, broad adoption of research data management in biological anthropology will hinge on a broad understanding of its benefits: new and more significant fields of research opened up by large bodies of reliable skeletal data.

The following sections elaborate on these aims.

2.1 Make Research Data Management a Central Concern in Research

Current approaches to standardised data acquisition in biological anthropology (e. g. Osteoware, AnthroBook, OsteoSurvey and CoRA; cf. section 1 of the project description) are focussed on application software supporting researchers during their investigations of human remains. The software solutions currently available and being developed help researchers to create datasets to be analysed, either on their own or, increasingly, in collaborative settings. Preparation of datasets for data pooling, publication, reuse or long-term storage has hardly been addressed so far[1]. But we believe these issues to become important in the near future

[1] But see session 17 at the 80th Annual Meeting of the American Association of Physical Anthropologists 2011 in Minneapolis (Minnesota, USA), entitled ”Data Management in the 21st Century: Integrating Bio- and Geo-informatics in Physical Anthropology”. Proceedings are published in the American Journal of Physical

for the following reasons:

1. Funding agencies increasingly demand data management plans and quality control for digital research data through their funding guidelines (National Research Council Committee on Responsibilities of Authorship in the Biological Sciences, 2003; Lämmerhirt, 2016).

2. Institutions curating digital research data are becoming aware of the challenges posed by securing and migrating their data holdings and are looking for sustainable solutions (Sholts et al., 2016).

3. Publishing primary data is going to become a measure for academic achievement, similar to conventional publications (Atici et al., 2013; Herold, 2015; Schiermeier, 2018).

4. Prehistoric and historic anthropology and their subdisciplines (e. g. paleopathology, paleodemography) leave a stage of methodological consolidation and will need to amass large bodies of high-quality research data in order to make valuable contributions to the understanding of human history (as demonstrated by the GHHP, cf. section 1 of the project description)(Sholts et al., 2016).

While these issues continue to become more and more pressing, researchers tend to avoid action towards improvement because this would entail extra work and there are no promising strategies for the development of good common practices. Still, it is necessary act before professional research data management becomes absolutely inevitable. The development of infrastructures is a constant process that requires both thinking ahead and improvement by trial and error. Infrastructures are developed now to be tested and ready when they will be urgently needed. Funding to build such infrastructures is currently available and should be employed. Still, the involvement of active researchers is essential to ensure that infrastructures meet the actual demands of the research community and will be broadly adopted. Not all researchers need to work on the development of research data management strategies but all should think about it.

We, therefore, intend to foster the coordination of initiatives for research data management in biological anthropology (cf. section 1 of the project description) and to create exposure for these processes in the scientific community. Research data management should become a

Anthropology, Volume 144, Issue S52.

regular topic of scientific conferences, publications and institutional management. This will help to identify requirements and communicate approaches towards solutions.

The professional discourse should not be dominated by the factors urging the introduction of research data management listed above but by the benefits for research itself:

1. Unambiguous definition of methods and procedures
2. Transparency of research data, their production, precision and options for quality control
3. Pooling of research data
4. Unobstructed exchange of research data between computer systems
5. Provision of research data to larger groups of researchers

Our project promotes semantic research data modelling as an improvement to research data management in biological anthropology. We believe that this technology has numerous advantages for the attainment of various tasks in this area. But it is not necessarily the single answer to all problems. An intelligent mix of various suitable technologies offers more flexibility and is more robust in the long run than monolithic solutions. We acknowledge contributions from other approaches and believe in good coordination among publicly funded projects to avoid redundant development work and ensure compatibility between systems. As infrastructure projects are currently not well represented, developments in other projects are easily missed. During the HSC project, our project group has become well connected and attained a position to bring similar-minded groups together. A mix of well-connected small-scale infrastructures will be more sustainable in the long run than monolithic solutions. This attitude is shared by the other infrastructure projects in biological anthropology.

2.2 Establish Semantic Research Data Modelling

The HSC project has identified RDF modelling of research data to have the potential to significantly advance research data management in biological anthropology. The key to this potential is the explicit formulation of the data's inner coherences. With RDFBones, for example, separate documentation of datasets is largely obsolete as the datasets are self-explaining, having the following types of information engrained with the actual research data:

Research design The outline of the scientific investigation producing the data is explicitly described, including the choice of material, methods to be employed, protocols to be adhered to and analytical strategies.

Line of reasoning The entire process chain of a scientific investigation is documented, explicitly stating which material was selected for a study, on what material a certain observation was made or analysis performed, how the resulting data were processed and on which data conclusions from the investigation are based.

Provenance Data include information on who funded and commissioned a scientific investigation, who developed the research design and who carried out the research.

Inclusion of additional information can be implemented in the future, e. g. on licensing, dataset versioning, research context or the intended archival time. Enriching research data in this way offers a number of advantages in comparison to conventional publications or even tabular data:

Transparency Researchers can assess data for plausibility and compliance with best practices before reusing them. Quality can additionally be ensured by selective re-examination of material.

Data reuse Researchers can reuse all intermediate products of investigations, not just the final results. These are are still fully documented.

Engagement Provenance information enables researchers to find contact partners in order to discuss research data and possibly avoid misunderstandings. It also helps to document diverging assessments made on the same factual basis.

A major advantage of data modelling is that semantic relations are maintained, even if data items are used selectively or pooled from various sources. This frees researchers from tracking and documenting data provenance when working with large amounts of data.

These advantages only play out if research data modelling is implemented in various areas of anthropological research. First and foremost, research data need to be coded properly and conveniently provided for reuse. But these efforts are worthless if researchers will not understand the potential of semantic research data repositories and conduct studies drawing on their potential. An introduction strategy will have to balance the creation of data repositories against demonstrating practical applicability in research. Both fields will need to grow alongside while stimulating each other. The establishment of semantic research data modelling hinges on success in raising awareness for the necessity of research data management (cf. section 2.1) as efforts for data enrichment and repository maintenance will only be made if the problems this solves are properly understood.

2.3 Establish Curation of Domain-specific Data Standards

Currently, research and data standards in biological anthropology are published without being backed up by consortiums guarding their further development and improvement. This has negative consequences, compromising the intensive work spent on their formulation. First, flaws that only become apparent when standards are regularly applied are not corrected on the level of the standards but in every single research project. This causes research data to be unnecessarily heterogeneous. Second, standards quickly loose touch with methodological developments and new research topics emerging in biological anthropology. They become outdated which reduces their application in research.

Instead, standards should have permanent maintainers, as it is common practice in other scientific disciplines. The tasks of such maintainers involve collecting error reports and feature requests and publishing improvements as a continuous series of versioned, backwards-compatible releases. Depending on the scope and popularity of standards, maintainers can be individuals or consortiums of researchers that meet regularly.

Maintainers are generally volunteers and their appointment is easier if researchers' work is heavily dependant on the development of a standard. Therefore, standards should be domain-specific with a clearly limited scope. With a very broad standard, covering all areas of research in biological anthropology, maintainers will inevitably have to deal with issues that they find difficult to decide and which affect their work only marginally.

One reason why standards are not curated in biological anthropology is that they are formulated as advice for best practise but are not binding in any concrete context. For example, as there are no formally maintained data repositories, there is no pressing need to confirm with their data standards. Therefore, the establishment of proper maintenance structures for standards will depend on the advancement of research data infrastructures and a growing general awareness for research data management among biological anthropologists (cf. section 2.1 and 2.2).

3 Strategy for Achievement of Aims

Recently, the German Council for Scientific Information Infrastructures (Rat für Informationsinfrastrukturen, RfII[1]) has summarised the situation of research data management in Germany and abroad (RfII – German Council for Scientific Information Infrastructures 2016, 2017) and formulated suggestions for improvement. We have used this information to set up a project strategy in three stages (not to be confused with the phases set out in the work programme of the project description) for achieving the aims related in chapter 2. At the moment, a concrete definition of the later stages seems not to be adequate as the situation of research data management is changing on all levels (RfII – German Council for Scientific Information Infrastructures, 2017, 31). On a European level, for example, the vision of a European Open Science Cloud (EOSC[2]) is taking shape, while the RfII is calling for a national research data infrastructure (Nationale Forschungsdateninfrastruktur, NFDI; RfII – German Council for Scientific Information Infrastructures 2016, 2) and Universities are developing their own strategies (see Wehrle et al. 2017 for Freiburg University). Additional initiatives exist on federal, national and international levels. In this situation, the RfII proposes a two-way strategy with scientific communities teaming up with infrastructure providers to meet nationally coordinated tasks (RfII – German Council for Scientific Information Infrastructures, 2017, 27)). We understand our project as a contribution of the bottom-up component of this scheme.

Our strategy is based on the principle to bring the RDFBones standard into practical application in research as soon as possible. This priority is chosen for several reasons. Practical application in real research scenarios is the best environment to identify flaws in the RDFBones concept and the requirements of researchers and research projects at an early stage. These issues should be settled before infrastructures become consolidated later on. Also,

[1] http://www.rfii.de; last accessed on 15 August 2018.

[2] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud; last accessed on 15 August 2018.

research projects create exposure by actively engaging researchers and through project presentations and publications. They also demonstrate the feasibility of the approach. Each early adopter successfully employing semantic research data modelling in their projects will act as a motivation for further projects to follow suite. All these advantages would be missed if initial emphasis was put on the creation of formal infrastructures. The RfII also sees user integration as a critical factor for success of research infrastructures (RfII – German Council for Scientific Information Infrastructures (2017, 16/17, 27)).

Later stages will have to focus on consolidating the technologies proved successful in research. Here, the focus will be on provision and curation of data, standards and research tools. The RfII observes that in many countries the maintenance of digital infrastructures is often organised as paid services offered both by academic institutions and commercial enterprises – a solution that is generally less favoured in Germany (RfII – German Council for Scientific Information Infrastructures 2017, 17–19, 30). How funding of infrastructures will eventually be organised is unclear today. With our funding proposal we try to initiate a process that can be continued in many, and possibly in several, directions. RDFBones is an open standard than can be served with simple tools like text editors and implemented into complex software applications at the same time. With the proposed creation of the AnthroGraph Software (see attachment 05 of the proposal) we try to provide a prototype that can serve both as a tool for apt researchers and a basis for professional software. Which line(s) of development are most promising is one of the things to be found out during the proposed workshop.

Project stage 1 will act as a teaser for semantic research data modelling in biological anthropology. Successful case studies employing this technology will demonstrate its capabilities and make its further potentials understood. The vehicle for creating these case studies is the software AnthroGraph. The aim of this stage is to create awareness for the necessity to create better research data and make semantic research data modelling generally known as a possible solution.

Project stage 2 will help to consolidate the infrastructures that developed in stage 1. We intend to win research institutions with strong infrastructures as partners to maintain instances of the AnthroGraph software on a permanent basis and provide its usage to either all researchers or to the ones affiliated with the providing institutions.

Project stage 3 will build up infrastructures that serve as general supply units, providing research data to all researchers in biological anthropology.

The three stages will be realised overlapping each other. All of them will initially need infrastructure funding but are intended to shift financing of infrastructure maintenance to research and institutional funds.

3.1 Stage 1: Initiation

The HSC project revealed the perceived need of institutions to do something about their research data without having a concrete concept of what this should be. We intend to discuss the problem with interested agencies and put out semantic research data modelling as possible solution. A crucial first step is to create successful use cases as an example for other institutions and researchers to get a better understanding of this solution. To put these measures in perspective, we intend to raise the topic of research data management with the scientific community and initiate discussion.

Stage 1 is intended to make semantic research data modelling known in biological anthropology and to clarify its relation to other approaches. It also proves the potential of semantic research data modelling to be realisable.

3.1.1 Coordination of Efforts and Public Representation

During the HSC project, contacts with other projects working on improvements in research data management in biological anthropology have been established which are still being maintained. On several occasions, the intention to initiate regular exchange was professed but, so far, no initiative has been taken. We believe our project group to be among the best connected in this research area and would like to use a continuative project as a focus point for the establishment of more formalised structures of exchange.

At the 2017 meeting of the Society for Anthropology (Gesellschaft für Anthropologie, GfA), the possibility of founding a work group for data standardisation and modelling was discussed with interested colleagues. Finding ten supporters for a motion (as required by the GfA statute) to establish such a group turned out to be unproblematic. We are currently working on a group description and a work programme in order to obtain the ten signatures and propose the foundation of the work group to the GfA members' meeting in 2019. The work group is intended to unite researchers who engage in the formulation of standards and the development of software tools. Work group meetings are intended to deal with general topics in data standardisation and modelling that are common to many projects and to give members the opportunity to inform each other about their work.

To initiate coordination on an international level, our funding proposal includes the concept for a workshop bringing together anthropologists working in projects related to research data management and German specialists for research data infrastructures (see attachment 07 of the funding proposal). The format complies with the combined bottom-up and top-down approach of the RfII (see above). Among the intended outputs of the workshop is a paper to be published in a major anthropological journal. This would be the first full-fledged scientific publication addressing this topic and is intended to promote research data management as a major topic. The workshop offers an opportunity to also discuss how to maintain coordination of efforts on a regular basis.

After the two successful sessions at the annual meeting of the American Association of Physical Anthropologists (AAPA, cf attachment 03 of the proposal) in 2017 and the American Academy of Forensic Sciences (AAFS, cf. attachment 04 of the proposal) in 2018, a possible format for regular exchange might be an annual session at a major conference. To maximise impact, the session could be organised on different meetings in turn. In addition to the AAPA and the AAFS, a further qualifying association might be the American Anthropological Association (AAA). Regular conference sessions will only make sense if colleagues who are active in research data management approve of the idea. The workshop will provide a good opportunity to discuss this and possible alternative options like mailing lists or video conferences.

3.1.2 Use Cases

As exemplary use cases we are looking for small research projects or scientific collections with a manageable set of requirements. These need to provide financial means to develop and maintain digital infrastructures. Use cases should be organised by respected researchers and institutions in order to create exposure within the scientific community.

To make semantic research data modelling applicable in research settings, automation of data enrichment is essential. The effort required by researchers needs to be minimised in order to make the approach attractive. The solution proposed by us is the creation of AnthroGraph, a versatile software that can be configured and extended to serve in various research settings (see attachment 05 of the proposal). AnthroGraph allows for the creation of rich datasets without the need to teach all collaborators the principles of semantic research data modelling. While the main objective of RDFBones is to create better research data, AnthroGraph will also have to stand up as a research tool to be used during investigations. This is essential to gather information on the entire process chain in an investigation and to secure a reasonable usability for being accepted by researchers.

Concept and development of the core AnthroGraph application will need to be funded through infrastructure programmes. This will be done in cooperation with the project group of the first use case to ensure software applicability. Adaption of the core application to the needs of the partner project will have to be undertaken there. With the first use case, creation of RDFBones extensions for use in the partner project will be done by our work group in order to produce prototypes of well-formed extensions as examples for subsequent project groups using the software. Further use cases will provide manpower to realise software adaptation and extension production. Our project group intends to provide professional advice and engage with these projects for the generation of feedback. Gradually, new adopters will have to meet these challenges on their own as documentation of resources improves and the number of researchers experienced with semantic research data modelling increases.

Part of the role model brought forth by the first project cooperation is the practice to publish RDFBones extensions in suitable repositories and to appoint maintainers for these modules. Structures need to be established that check back on these maintainers in regular intervals to maintain an overview of active and abandoned extension projects. This task could be taken over by the GfA workgroup for data standardisation and modelling to be established (see section 3.1.1).

For further use cases it is desirable to demonstrate the capabilities of semantic research data modelling for interdisciplinary research (e. g. in archaeology, medicine or forensics). These cooperations will require project partners to make a strong commitment by modelling the non-anthropological data they provide. Such cases would demonstrate the high versatility of the approach.

New adopters of AnthroGraph will need to invest in adapting the software to their needs (see above). This will also contribute to the software's development and maintenance, thereby keeping the project alive during its initial stage. A stable future of AnthroGraph can only be secured by long-term commitment of institutions using it. The proposed workshop is intended to provide perspectives for such an institutionalisation.

3.2 Stage 2: Consolidation

Once AnthroGraph has been tested and optimised through deployment with initial use cases, it is intended to provide the software to researchers on a more permanent basis. AnthroGraph should become a tool that is not just employed on the basis of individual projects but can be used by researchers for all their investigations. To this end, we intend to cooperate with large research institutions who will provide AnthroGraph as a central resource for their research groups and partners. This will increase availability among researchers and initialise the formation of larger bodies of compatible data.

Possible candidates to maintain deployments of AnthroGraph are research collections of the magnitude of the State Collection of Anthropology and Paleoanatomy Munich (Staatssammlung für Anthropologie und Paläoanatomie München, SAPM) or the Natural History Museum in Vienna that provide sufficient staff and technical infrastructure, research institutions like the Senckenberg Nature Research Society or academic IT services with a strong research data management infrastructures. Partner institutions need to enjoy broad confidence among researchers who are asked to entrust them with their research data.

AnthroGraph is designed as a web application in order to serve large numbers of users from a single installation, thereby minimising efforts for software maintenance. This advantage plays out best with large user bases. Continued software administration by institutions also provides a backbone for maintenance of the software in an open-source project.

Another area that needs to be consolidated in project stage 2 is the coordination of maintained extension projects. There should be registries of active projects to help researchers find suitable extensions for their research, focus extension development and avoid redundancies. Also, systems of quality control need to be established, assessing both scientific relevance (e. g. peer review) and technical integrity (e. g. benchmarking). Who will be in a position to form such infrastructures and how they are to be organised will have to be decided at a later stage, based in developments in biological anthropology during project stage 1 and in research management at large.

3.3 Stage 3: Overarching Infrastructures

Once larger amounts of research data will have accumulated, they need to be made available for scrutiny and reuse. This is the ultimate purpose establishing semantic research data modelling without which work during the previous stages would be dispensable. Large distribution systems will be needed, offering researchers the opportunity to publish primary data and serving the international scientific community. Provision of research data will have to acknowledge ethical and legal constraints that might exist for certain usages of particular data. Provision systems might evaluate such constraints themselves or be limited to establishing contact between prospective users and providers of data. Due to the common RDFBones data model, datasets obtained from different sources can be easily pooled.

RDF data are boundless in nature, i. e. they can be extended in all directions and can form complex knowledge graphs on diverse contents. The formation of standardised datasets is not an intuitive outcome of semantic research data modelling. Large knowledge graphs of osteological information can be envisaged as scientific resources. Provenance would be intrinsically documented in these networks, facilitating acknowledgement of providers. A difficulty in providing such information as open data, however, would be the enforcement of usage restrictions (see above). A possible solution would be closed knowledge graphs to which queries can be sent that are object to scrutiny by curators or for which only queries for summarising results are eligible.

How provision of research data will be organised is impossible to foresee now, given the developments in research data management on various levels (see above). It has to be assumed that generic systems for research data provision will emerge that will just require biological anthropologists to define data models and search keywords. A front office - back office concept as exemplified in the Netherlands (RfII – German Council for Scientific Information Infrastructures 2017, 17, 29/30) is also feasible. A further development of AnthroGraph into a complex data distribution platform is possible but most likely not a reasonable solution for a relatively small community like biological anthropology. The overall aim of our project is achieved by the existence of high-quality research data that will be of value to future generations of biological anthropologists.

Bibliography

Atici, L., Kansa, S. W., Lev-Tov, J., and Kansa, E. C. (2013). Other people's data: A demonstration of the imperative of publishing primary data. *Journal of Archaeological Method and Theory*, 20(4):663–681.

Herold, P. (2015). Data sharing among ecology, evolution, and natural resources scientists: An analysis of selected publications. *Journal of Librarianship and Scholarly Communication*, 3(2):1–23.

Lämmerhirt, D. (2016). Disciplinary differences in opening research data. PASTEUR40A Briefing Paper.

National Research Council Committee on Responsibilities of Authorship in the Biological Sciences (2003). *Sharing Publication-related Data and Materials: Responsibilities of Authorship in the Life Sciences.* National Academies Press (US), Washington D.C.

RfII – German Council for Scientific Information Infrastructures (2016). Enhancing research data management: Performance through diversity. recommendations regarding structures, processes, and financing for research data management in Germany. Rfii recommendations, RfII – German Council for Scientific Information Infrastructures, Göttingen.

RfII – German Council for Scientific Information Infrastructures (2017). An international comparison of the development of research data infrastructures: Report and suggestions. techreport 5, RfII – German Council for Scientific Information Infrastructures, Göttingen.

Schiermeier, Q. (2018). For the record: Making project data freely available is vital for open science. *Nature*, 555:403–405.

Sholts, S. B., Bell, J. A., and Rick, T. C. (2016). Ecce homo: Science and society need anthropological collections. *Trends in Ecology & Evolution*, in press:in press.

Wehrle, D., Wiebelt, B., and Suchodoletz, D. v. (2017). Design eines FDM-fähigen Speichersystems. In *10. DFN-Forum Kommunikationstechnologien, 30.-31. Mai 2017, Berlin, Gesellschaft für Informatik eV (GI)*, pages 115–124, Berlin.