Scope of AnthroGraph

Requirements for the Development of a Minimum Viable Product

Felix Engel, Stefan Schlager

17th August 2018

Felix.Engel@anthropologie.uni-freiburg.de

Contents

Nomenclature

AnthroBook	Data acquisition software developed by the SAPM
AnthroGraph	Software to be developed during the proposed project
AnthroGraph MVP	Nuclear version of the AnthroGraph software to be produced with the proposed funding
API	Application Programming Interface
AQUiLA	Information system developed and maintained by the Senckenberg Foundation
CoRA	Commingled Remains Analytics
CT	Computer Tomography
DFG	Deutsche Forschungsgemeinschaft (German Research Foundation)
FACTS	Forensic Anthropology Center of Texas State University
GUI	Graphical User Interface
HSC	Human Skeletal Collections; research project during which the digital data standard RDFBones was developed
MVP	Minimum Viable Product
OBI	Ontology for Biomedical Investigations
OsteoSurvey	Data acquisition software for mobile devices

Osteoware	Data acquisition software developed and provided by the Smithsonian Institution (Washington, USA)
PDF	Portable Document Format
Protégé	Popular RDF editor
RDF	Resource Description Framework
RDFBones	Digital standard for osteological research data developed during the HSC project
SAPM	Staatssammlung für Anthropologie und Paläoanatomie München (State Collection for Anthropology and Paleoanatomy Munich)
SPARQL	SPARQL Protocol and RDF Query Language
URI	Uniform Resource Identifier
URL	Uniform Resource Locator

1 About this Document

This document is about the development of a software which here is referred to by the working title ”AnthroGraph”. AnthroGraph is a server application that helps scientific institutions to set up information systems for managing research data emanating from investigations human remains and from research in biological anthropology in general. It implements the digital standard RDFBones for research data from osteological investigations which was developed during the research project ”Human Skeletal Collections” (HSC), funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG).

The HSC work group is applying for funding of the development of AnthroGraph to the DFG. This document is part of the funding proposal which is submitted within the DFG funding scheme ”e-Research Technologies”. Knowledge of the funding proposal is assumed as it provides background information that has a bearing on the scope of the software discussed here.

1.1 Aim and Scope of this Document

The objective of this document is to define the scope of requirements for an initial realisation of AnthroGraph as a minimum viable product (MVP), here referred to as ”AnthroGraph MVP”. In this context, the term MVP refers to the most simple application design that can be successfully employed in a realistic research scenario. In practice, AnthroGraph MVP will serve certain use cases but will not support the full scope of use cases that can be envisaged for AnthroGraph in the medium and long term.

While this document’s primary objective is to define the scope of AnthroGraph MVP, some information on the larger concept for AnthroGraph is also related in order to convey an idea of how this initial development stage is framed. In the following, sections entitled ”Perspectives for Further Development” contain possible outlooks onto future stages of the Anthro-

Graph software that should be considered while realising the MVP.

1.2 Function Within the ”e-Research Technologies” Funding Proposal

The requirements described in this document were compiled in the course of the HSC project. In this context, representatives of collections holding skeletal human remains were interviewed and completed a structured questionnaire. A two-days workshop with biological anthropologists collected their requirements for and expectations towards digital research tools. Representatives of the HSC project participated in two international symposiums on data standardisation and management in biological anthropology to discuss approaches and technologies with other researchers that are active in this domain.

These insights led to the formulation of the grant proposal within the ”e-Research Technologies” funding programme. The development of AnthroGraph is proposed there as an initial step in a broader strategy to introduce semantic research data modelling in biological anthropology.

2 Purpose

2.1 General Purpose

AnthroGraph is a tool for researchers in biological anthropology that enables them to code the data resulting from their investigations into RDF (Resource Description Framework) graphs without requiring advanced knowledge of this technology. By virtue of RDF data modelling, these data can be transformed according to various data models to produce required input for other software tools or to be archived as highly annotated and self-explaining datasets. While data input can be performed by persons of moderate computer literacy, power users can configure which kinds of data are supported. In this way AnthroGraph supports researchers in the development of improved or novel study designs.

AnthroGraph is conceptualised as a software for collaborative research in specific research contexts, e. g. within a research project or institution. It can be customised to suit the specific requirements of such use cases. Institutions can host the software on their own network servers to provide its functionalities as a service to a specific target group of researchers.

By default, the software is designed for documenting research on skeletal material, as this represents the bulk of material researched in biological anthropology. However, it can be configured to cover other types of human remains as well. These are not implemented as standard features because their variability and rare occurrence do not merit the effort to cover all kinds of preservation and preparation.

2.2 Field of Application

AnthroGraph has the following functionalities:

1. Manage collections of human remains

a) by issuing URIs for collection objects

b) by representing systems of ordering

c) by supporting the compilation of human remains inventories

d) by documenting events in the material's curation

e) by making information on the collection public

2. Support and document research on human remains

   a) by representing research projects

   b) by supporting various types of investigations

   c) by representing contextual information

3. Manage digital assets relating to human remains

   a) by providing metadata schemes for different types of media

      i. images

      ii. 3D representations

      iii. text documents

   b) by documenting their relation to research objects and data items

4. Enhance existing research data on human remains

   a) by explicit specification of study designs

   b) by annotation

5. Transform research data to comply with specified data models

   a) to pool research data from disparate data sources

   b) to analyse data according to the logic of the research objective

   c) to provide input for other software

6. Perform custom queries on the data

A central principle of the RDF is reference of concepts by uniform resource identifiers (URIs). With human remains to be used in scientific investigations, URIs should be issued by the respective institutions curating the material. Institutions holding collections of human remains can use AnthroGraph to define URIs for their collection objects and provide them along with related information for reference in research. This should contain information on where specific collection objects are kept in storage facilities in order to facilitate physical re-examination and under which identifiers they might have been referenced in other contexts. AnthroGraph supports definition of multiple systems of ordering like shelf or archive numbers.

Inventories of human remains in collections record their completeness and preservation status at a given point in time. A series of inventories taken at different times can monitor changes in preservation and therefore be used for quality assurance in collection management. But inventories are also typical starting points for investigations, evaluating the expected scope of information to be gained from human remains. The completeness of collection material is influenced by loss and destruction, e. g. caused by probing for destructive analytical methods. AnthroGraph records such occurrences as curation events in a collection's history. Other curation events can be restorative measures, changes in storage conditions and similar occurrences.

OteoGraph helps research collections to make selected information about their holdings public. Examples of such information are a collection catalogue listing skeletal series and completeness information or a list of available research data. Publishing such information might help researchers who are interested in using material from a collection to plan their requests more precisely before contacting curators, reducing their workload.

With AnthroGraph, it is possible to register research projects and provide related information, e. g. about the composition of the work group, objectives, the project's concept or funding. This allows to understand the context in which research data were produced. Projects contain a variable number of investigations on human remains that can be of different types (e. g. age and sex estimation, stature estimation, investigations of pathological traces).

While all investigations follow an identical overall procedure predetermined by the Ontology for Biomedical Investigations (OBI) containing steps like specimen collection, assays, data transformations and conclusions, individual investigation types differ from each other in their specifications, e. g. types of material input, assays to be conducted or data output. Each investigation type is based on a particular study design, providing detailed documen-

tation of its specified components. Specifications for investigation types can be written by users and loaded into the AnthroGraph software to complement the software's scope. This mechanism is the core concept of AnthroGraph.

AnthroGraph can be configured to represent contextual information (e. g. archaeological excavation units, circumstantial evidence from forensic cases, taphonomic environment data) needed for analysing the research data produced from the human remains. This allows for employing the software in various research contexts.

AnthroGraph provides a data store where files can be managed and searched for. In addition to a generic metadata scheme, additional sets of metadata are provided for images (including x-ray and output of other imaging techniques), three-dimensional representations (e. g. 3D scans or CT images) and text documents. 3D objects may replace actual human remains as specimen for investigations (virtual anthropology). Files can be used as digital assets for collection curation, research projects and investigations. Their relevance to specific elements of these processes can be specified with semantic relations.

Apart from supporting structured data entry, AnthroGraph provides import routines for external, tabular, data during which users specify the semantic relations between data items. This process defines concrete study designs for previously undocumented datasets. Information density can be further improved with annotations on how data was produced, e. g. names of researchers who performed individual investigations or their relations to research projects.

Importing external data automatically transforms their data model from the one of the originating database to the one defined by RDFBones. In this way AnthroGraph pools data from disparate datasets and makes them directly comparable, a feature valuable for research projects working on large amounts of data from various sources. But transformation of data models can also be used to other ends. Data can be analysed according to a new data model with an internal logic that differs from the one of RDFBones. Also, output filters providing tabular data as input to other software tools (including custom software like R packages) can be defined and routinely employed.

AnthroGraph provides an interface for complex data queries based on the SPARQL Protocol and RDF Query Language (SPARQL).

AnthroGraph MVP supports the functionalities listed above to the degree demanded by the deployment scenario defined for this version (section 2.4).

2.3 Deployment Scenarios

AnthroGraph supports collaborative research of biological anthropologists working in a specific research context and should be deployed by an institution central to this context (cf. section 2.1). It is targeted at institutions that do one or several of the following:

1. Conduct a specific research project on human remains that produces research data
   a) collaborate with other institutions on such a research project
   b) open a collaborative research project to researchers for contribution
2. Conduct scientific investigations on human remains on a regular basis
   a) collect the research data emanating from such investigations
      i. produce standardised documentation of such investigations
3. Curate a collection of human remains and provide the material for scientific research
   a) collect the research data emanating from such research
4. Provide research data from various sources to researchers
   a) for reference
   b) for use in other research projects

The following sections give detailed descriptions of these scenarios.

2.3.1 Research Project

A team of researchers unites for the joint examination of a defined body of human remains. Individual researchers provide their specific expertise with certain types of investigations. On the other hand, several researchers collaborate on the same types of investigations, raising the issue of inter-observer errors (cf. figure 2.1 a). The leading institution among those supporting the project employs AnthroGraph to coordinate the project group and its contributions.

As the project members all work towards the same goal, mutual trust and interest in each other's work can be expected. Therefore, all users of the information system should be able to see their colleagues' work. In order to maintain work coordination, however, it might be

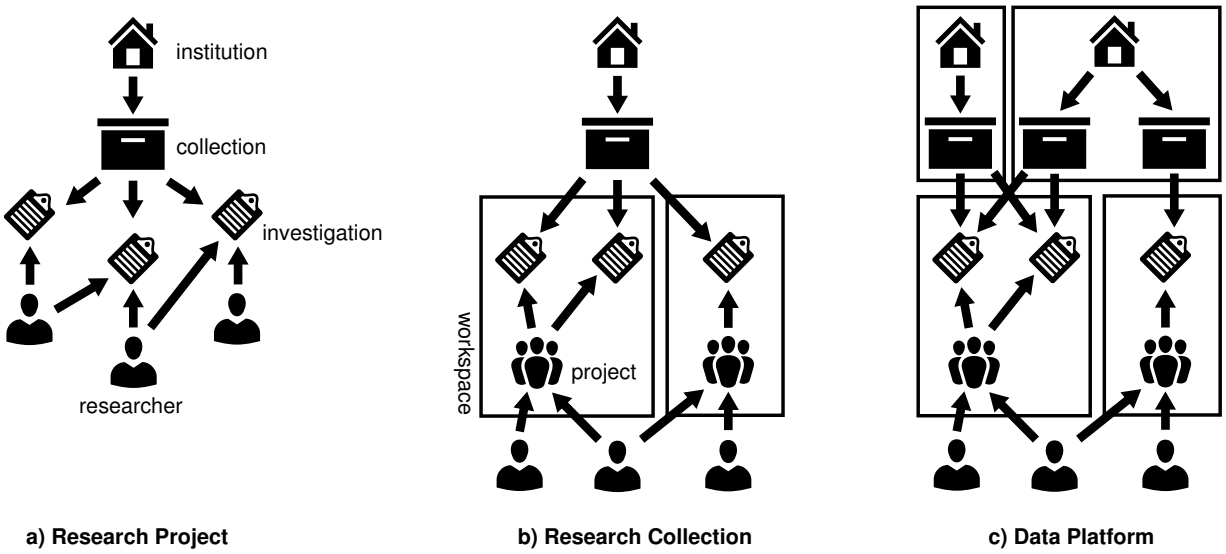

Figure 2.1: Schematic representations of deployment scenarios.

beneficial to restrict write permissions to change existing data to their respective authors and system administrators.

Part of the project's objectives might be the provision of project results to external long-term storage facilities or to make them publicly available through the information system itself.

An example for this scenario is the Phaleron Bioarchaeological Project.

2.3.2 Research Institution

The researchers working at an institution routinely conduct research projects involving human remains. The institution uses AnthroGraph as a research environment for these projects and for storing the resulting research data. Groups of assigned researchers work on the active projects while concluded projects are stored as archived bodies of data.

Depending on the institution's organisation and purpose, discretion between project groups might or might not be an issue. As a consequence, solutions might rather fall in line with the research project (section 2.3.1) or the research collection scenario (section 2.3.3). A strict separation of projects might be necessary with institutions handling sensitive data, e. g. in forensics or medicine. For other institutions it might suffice to remove write permissions for

concluded objects.

An example for this scenario is the Forensic Anthropology Center of Texas State University (FACTS).

2.3.3 Research Collection

An institution holding a research collection of human remains provides the material to researchers for their investigations. These agreements oblige the researchers to leave the research data that they produce with the curating institutions in order to promote reuse of existing data and minimise handling of the original material. With AnthroGraph, institutions can provide a research environment for scientists working with their material which they can configure to set their own requirements of data standardisation but also to accommodate the specific needs of each research project.

As the research projects are carried out by scientists from different institutions and research backgrounds, mutual trust between them can be neither expected nor demanded. As a consequence, researchers should not be able to access data outside their own projects. These have to be restrained to their own respective workspaces set by the software (figure 2.1 b). A digital rights management needs to allocate access rights by users to projects so that several researchers can collaborate on one project and researchers can be part of several projects.

The curating institution will need to query data from all projects, at least on the level of available investigation types, in order to check if novel applications for usage duplicate previous research conducted on the material.

An example for this scenario is the State Collection for Anthropology and Paleoanatomy Munich (Staatssammlung für Anthropologie und Paläoanatomie München, SAPM).

2.3.4 Data Platform

An institution uses AnthroGraph to maintain a data repository where several curation facilities of human remains collections provide information on their holdings and research data from investigations of these materials. Research projects use the software for researching suitable material and/or existing data in preparation for their projects, obtain agreements with the respective curators and conduct their investigations. Here, the information system acts as a broker of scientific information between institutions and researchers.

Both curating institutions providing information about their collections and researchers conducting their projects need shielded workspaces to manage their contributions apart from publishing some of the outcome on the information system (cf. figure 2.1 c). Depending on the sensitivity of the data exchanged, securing the separation of work spaces might be more important than with the research collection scenario (section 2.3.3).

Currently, a data platform as described above does not exist. Some aspects of this vision are realised in the AQUiLA system by the Senckenberg Foundation who provide the system to external research collections to manage their holdings and offer research functionalities across collections to researchers[1].

2.4 Purpose of AnthroGraph MVP

AnthroGraph MVP focuses on the simplest of the deployment scenarios described above, the research project (section 2.3.1; figure 2.1 a). The functionalities listed in section 2.2 are supported to the degree to which they are relevant for this use case.

Management of one body of skeletal material is supported, including definition of URIs, modelling systems of ordering, compiling inventories and maintaining a collection history through a series of curation events.

The software provides a generic workflow for the conduction of investigations, oriented at most common practices in biological anthropology. Which concrete types of information the software is able to process is determined by configuration and customisation of the system.

Digital assets can be stored in the application's file store or in a specified external directory. Besides a default set of metadata, the application provides a more specific metadata scheme for images. Assets can be linked to pieces of information by definition of semantic relations.

Data output can be configured by definition of parameter queries to be executed from relevant screens of the graphical user interface (GUI). This mechanism facilitates data extraction for users that do not know how to perform SPARQL queries. In addition, the software offers an interface or running custom SPARQL queries on the entire database.

[1] https://search.senckenberg.de/aquila-public-search/search; last accessed on 2 July 2018.

2.5 Perspectives for Further Development

AnthroGraph MVP provides a basis for also realising other deployment scenarios as described in section 2.3. To find out which of these are needed within the scientific community and what specific requirements this would entail is an objective of the proposed project.

Another trajectory for future diversification of the software are versions serving more specifically individual research contexts (e. g. forensics, bioarchaeology, repositories of 3D representations, medical applications). The versatility of the RDF allows for creating interdisciplinary information systems that extend the representation of contextual data into full-fledged support for related disciplines.

3 Design Principles

The following sections detail general characteristics that AnthroGraph needs to display in order to have the intended impact with the promotion of semantic research data modelling in biological anthropology. These principles guide all aspects of its development.

3.1 Implementation of RDFBones

The main incentive for creating AnthroGraph is to make the results from the HSC project available to the research community. Therefore, RDFBones is set as the application's data model.

RDFBones provides a core ontology containing general concepts for representing research in biological anthropology and a set of rules how to extend the ontology by creation of subclasses of the classes representing these concepts. This structure ensures that ontology extensions can introduce new concepts while maintaining the same overall structure.

For the implementation in AnthroGraph this means that workflows determining how users navigate through the application can be defined on the level of the core ontology to achieve a product that works for all extensions to be added in the future. It also means that SPARQL queries defined on core ontology level will work for all data acquired through ontology extensions of all kinds. Ongoing curation of the RDFBones data standard will involve concepts being continuously added to the core ontology in order to increase compatibility of data coded with RDFBones. For the development of AnthroGraph it should be considered how to prepare the application for upgrades to higher core ontology versions.

Ontology extensions can be used to enlarge the scope of the core ontology in the following areas:

- collection management
- human remains inventories

- research designs (types of investigations)
- contextual data
- mapping of external data models
- preparation of output (e. g. definition of data tables)

Each of these issues comes with a set of rules how extensions need to be written and which classes of the core ontology they can extend.

AnthroGraph MVP is designed to support three types of ontology extensions commonly needed for the realisation of research projects: skeletal inventories, research designs and contextual data.

3.2 Server-based Application

AnthroGraph is developed as a network-based application to be deployed on a local area network or as a web application on the internet. This architecture is preferred over client software solutions because it is independent of end users' operating systems and minimises maintenance efforts. Institutions concerned with data security may opt for an in-house deployment on a local network while others relying on external cooperation will prefer deployment as a web application. Users preferring a standalone application on their personal computers can deploy the software on localhost.

3.3 Adaptability

A key feature of AnthroGraph is its high customisability and adaptability both by institutions deploying the software and by researchers using the software. Institutions may want to add static pages to the GUI or implement their own workflows. Tool selection and software development should minimise the effort needed for such alterations and draw on technologies that can be employed by apt researchers rather than IT professionals where possible. Basic RDF knowledge empowers researchers to write their own ontology extensions giving them the opportunity to adapt their work environment to the needs of their research.

4 Types of Data

This section gives an overview what kinds of information AnthroGraph processes. Most of this information is structured and, therefore, open to digital data processing. Where unstructured information needs to be recorded, this is specifically pointed out.

4.1 Collection Objects

AnthroGraph is focussed on bodies of material of particular interest for the investigations carried out with the software, referred to as collections. What exactly constitutes collections has to be defined for each deployment scenario (section 2.3) or use case. AnthroGraph MVP provides for one collection but use cases involving several collections are feasible.

Collection objects are anticipated to consist of human remains and their physical and virtual representations. The bulk of material are skeletal remains but mummified remains, anatomical dry and wet specimen and other forms of preservation are also possible. Collections might contain replica of human remains or consist entirely of digital representations, e. g. computer tomography (CT) scans.

Collection objects are organised according to various systems of ordering, e. g. shelf, acquisition or archive numbers. The system of ordering that is currently in operation needs to be clearly identified but representation of depricated systems is also necessary as they might appear as references in older documents.

4.2 Human Remains Inventories

Inventories are datasets recording the completeness and preservation of a certain body of human remains. Completeness can refer both to individual bones or to expected sets of bones, most commonly skeletons. Information on the degree of preservation clarifies if a specific

observation is possible, for example if the bone surface is intact and available for inspection. Different study designs have their own specific requirements concerning material completeness and preservation. Therefore, a variety of inventory types is needed, each assessing material quality in different ways. Inventorying is a common first step in investigations of human remains to clarify the availability of material. But it is also performed in the context of curation, e. g. to monitor the quality of collection objects or to assess the need for conservation measures.

4.3 Archival Information

Professional collection management is not within the scope of AnthroGraph. However, some information from this domain has a bearing on if and what kind of research can be performed on human remains and should be recorded.

Depending on the provenance of human remains and the legal framework within which they are curated, access to the material might be restricted to certain groups of people. Also, the scope of investigations that can be legally performed on the material might be limited (e. g. general prohibition of invasive methods). Similar restrictions can be imposed on the grounds of the material's uniqueness or its state of preservation.

Assignment of such restrictions is a curation process that should be documented with a collection's history. Information on an object's current legal or curation status should also be readily available to support planning of studies. Archival information, however, is usually specific to individual collections and needs to be configured by their respective curators.

4.4 Project-related information

Research projects provide the framework in which investigations are performed. Information on their aims and objectives, their organisation and sources of funding help to understand why investigations were carried out and why they were carried out in a specific way. While much of this information is rather variable and needs to be provided in an unstructured form (possibly as a text document file), structured information includes the specification of project members, the capacities in which they contributed and the time spans of their involvements. Researchers reviewing or reusing project results will need this information for finding con-

tact persons able to provide information that had not been documented during a project's execution.

4.5 Investigation Data

Investigations are carried out in order to make structured observations on or conduct analyses of human remains. The products of these assays are pieces of recorded data. On their own, however, these data items are quite meaningless. More information needs to be gathered to explain why and how they were obtained in order to draw some meaningful information from them. In order to gain an explicit record of investigation data, all stages in its process chain need to be documented. This ensures transparency and provides ample opportunities for reuse, not just of the final results of an investigation but also of its intermediate products.

Which processes are carried out in an investigation and how they ought to be performed is documented in a study design. A study design states the investigation's overall objective while also specifying objectives for all processes in its chain of action and explains its investigative strategy by defining dependent and independent variables. It also contains specifications of the types of human remains needed, of all assays and data transformations involved and of all data items to be produced. Study designs document investigations on a generalised level without reference to concrete material or research contexts.

Information on a specific investigation is documented in an investigation plan. This outlines why a specific type of investigation is applied on a particular body of material to achieve a particular end. In most cases, this information is unstructured and possibly provided in the form of a text document file.

Crucial for the performance of an investigation is the decision on whether a certain material is suitable or not. Exclusion of material affects a study's outcome by lowering sample size and increasing the relative weight of evidence from other individuals. Therefore, specimen collection should be documented. It can involve assays of its own. Specimen collection can be limited to the selection of usable material but might also involve preparation processes (e. g. the creation of thin sections or extraction of certain tissues), possibly damaging or destroying parts of the material to be studied. In any case, there is a difference between the material considered for an investigation and the material on which the investigation is even-

tually performed, referred to as specimen. The nature of this difference in a specific type of investigation needs to be explained in its study design.

On the specimen in an investigation, one or several assays are performed, resulting in a series of data items. These can be measurements on some scale, categorical or textual in nature. Investigation designs might prescribe data transformations to be performed on the output of assays. Such transformations can either alter individual data items (e. g. by multiplication with some factor) or merge several items into one value (e. g. by calculating an arithmetic mean). Several such transformations are possible, both in parallel and in sequence.

From the data generated in the course of an investigation, conclusions are drawn. These are the final results of an investigation and may involve expert judgement by executing researchers. An investigation can have several conclusions.

Finally, AnthroGraph can record which parts of an investigation are explained in a certain document and which results contribute to a certain publication. Such connections help a better understanding both of how investigations were conducted and of the resulting documents.

4.6 Additional Information

In most realistic research scenarios in biological anthropology, investigation results are not self-sufficient but need to be set against some external source of information to assess their biological or social significance. The concrete nature of such external data is highly variable and cannot be foreseen in the development of AnthroGraph. Instead, the software needs to be customisable to integrate external data structures. Another incentive to integrate external data are investigations working with documented data from past investigations that were not conducted with AnthroGraph. These can be pooled with data from within the information system, have data transformations performed on and conclusions drawn from them.

To document the origin of published data, it is necessary to unambiguously identify the source by sufficient bibliographic information. Suitable external information (e. g. archaeological excavation units) can be represented as systems of ordering. AnthroGraph should support representation of geographical data for spatial analyses. Other data can be simply defined as data items. As they exclusively serve as contextual data, their integration into a semantic framework is optional.

4.7 Annotations

No data acquisition scheme is perfect and cases of uncertainty are always to be anticipated. While decisions have to be made how to express information in the framework of a specific scheme, problems with assignment need to be documented. Therefore, users can leave comments with data entries of any kind. Comments can exclusively refer to one information item or relate several items with each other. They can also incorporate citations of passages from written documents.

5 Workflow

This section outlines common research routines in biological anthropology that AnthroGraph needs to support and points out important aspects in their implementation. figure 5.1 summarises application components suggested by workflows and sketches likely navigation behaviour between them.

5.1 Collection Management

5.1.1 General Information

Management of general information on collections involves entering a textual description, specify the institutions sustaining the collection and define collaborators who can be entered with their roles and durations of service.

5.1.2 Information on Curation

Changes in conditions under which a collection is kept or in how the collection is managed need to be recorded and the information made available to researchers working with collection objects. Examples for such information are access restrictions to the material, guidelines for its handling or equivalence of URIs. RDFBones organises this kind of information as curation events. These can refer to the collection as a whole or to individual collection objects. Their duration is variable, open-ended events indicate current status information.

Management of information relating to curation involves an option to create and edit curation events. Objects that an event refers to can be selected by collection, identifier or object URIs.

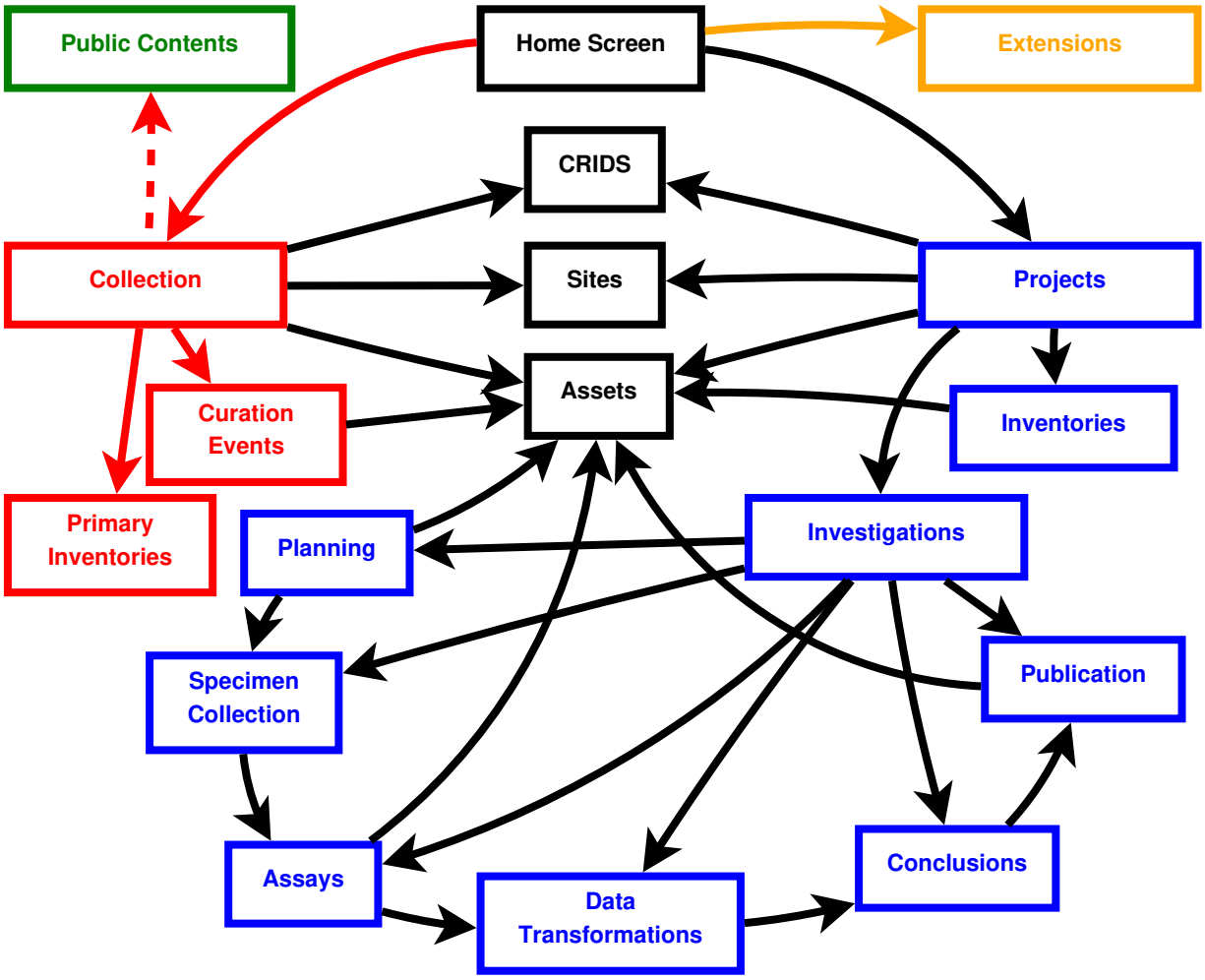

Figure 5.1: Sketch of application components in AnthroGraph and likely navigation paths (solid arrows). Role assignment govern write permission for blue and read/write permissions for red elements. The dashed arrow represents configuration of public data display. Orange elements are exclusively accessible to system administrators.

5.1.3 Primary Inventory

A central task of collection management is to register the human remains from the collection (cf. section 4.1) with the information system. It is important that this is performed by an authoritative group, the collection managers (or curators) to avoid the creation of duplicates. Investigations can exclusively reference material that has been entered by collection managers first. For this purpose, RDFBones provides a very simple form of inventory (primary inventory). This is based on natural anatomical entities (e. g. individual bones or teeth) and exclusively reports if these are complete or only preserved in part. In cases where two independently registered fragments turn out to belong to the same element, the two URIs are marked as equivalents by the respective RDF statement (owl:sameAs). Such statements have grave bearing on analyses of the material and should only be made after careful examination and verification of the underlying facts. If two URIs are found to refer to the same element, the statement of their equivalence is recorded as a curation event with semantic relations to both URIs, justifying the measure and describing in detail how to identify the two fragments. Researchers can continue to reference the two initial URIs while documenting investigations, depending on which fragment observations are made on.

The URIs assigned to human remains during primary inventory should be truly unique and function as worldwide identifiers for the material they represent. RDF URIs are commonly URLs (uniform resource locators) and AnthroGraph should provide a possibility to configure the domain in which URIs are created.

Individual primary inventories represent bodies of material that are meaningful to the organisation of the collection. What this means for a specific collection depends on its nature and context, common examples being forensic cases, archaeological excavation units or medical examinations. The system of ordering according to which these units are organised is referred to as the collection's "primary registry". Primary registries should represent the acquisition of material by collections rather than information derived from it (e. g. assignment to individual skeletons) as their structure needs to remain unchanged. They should also be in regular use within the institutions curating the collections and ideally being a key to identifying material in these institutions' holdings.

5.1.4 Systems of Ordering

Long-established collections might have several systems of ordering for the objects they contain, both historical and current, referencing the material in diverging ways. All these systems are maintained in collection management as registries but only the primary registry can be used to register actually existing human remains. Ordinary collection registries can reference material already registered through primary inventories or record material that since has been lost.

Management of registries involves an option to enter general information and to create and edit identifiers within the registry. Identifiers can have several meaningful components and reference various objects. Search functionalities are needed to find identifiers referring to certain types of material or identifiers referring to the same object.

5.2 Creation of Inventories

Inventories (4.2) need to reference objects that have already been registered with the system through primary inventories (5.1.3). They can refer to particular features or define arbitrary sections of these objects and contain all kinds of data items related to their completeness and preservation.

Implementation of inventories in AnthroGraph needs to meet the challenge of minimising the effort users have to spend on data entry. It should be possible to assign information on completeness and preservation to every single anatomical element (e. g. bone) and even to specified parts of these (region of interest). Alternatively, users should be able to assign this information to entire anatomical contexts (e. g. the cranium or the dentition) if it is identical for all its elements. In this case, the software is expected to automatically create individual assignments for all elements in the specified context which are defined by the inventory scheme.

Inventories of the same kind can build up on each other. This is the case if an investigation is performed on the same material as a previous one and researchers want to check for changes since the last inventory. To this end, it should be possible to load the previous inventory to just alter the data entries that have changed.

Inventories may have digital assets (e. g. homunculus diagrams, photographs).

5.3 Project Management

General information about projects includes a textual description and specification of participating institutions and people, both with the roles in which they contribute. There needs to be an overview of investigations carried out in the course of the project along with a functionality to filter them by investigation type. A similar overview should present publications of project outcomes.

Another concern of project management is the management of skeletal inventories to provide a basis for its investigations. Inventories need to provide information required by the investigations' study designs. Users either select existing inventories or create new ones. With the latter option, it needs to be possible to load an existing inventory as a basis for an updated version.

As an alternative or addition to the management of inventories, previously documented investigations can be selected as input for a project's investigations.

5.4 Execution of Investigations

For the documentation of investigations, AnthroGraph users can select among the types loaded into the software as extensions. RDFBones describes investigations as a succession of the following processes:

1. planning
2. specimen collection
3. execution of assays
4. data transformations
5. drawing conclusions
6. publication

A generic GUI, designed to support all kinds of extensions to be loaded, needs to follow this outline, providing input screens for each process. Many projects study series of human remains, e. g. a number of skeletons from a cemetery. In such scenarios, two basic strategies

are feasible: either researchers complete the entire investigations for each individual, one after the other, or they perform one examination step on all individuals before they move on to the next. For example, it might make sense to prepare all specimens as thin sections before conducting assays on a microscope. In order to support both strategies, AnthroGraph needs to provide two trajectories of navigation, one going through all investigation stages for one individual, the other going through all individuals for one stage. Offering both directions of navigation, towards the next stage or towards the next individual, on every input screen provides most flexibility.

The planning stage involves specifications on which material is investigated and with what intent. An investigation type's study design defines which kinds of human remains and which information on their completeness and preservation are needed as input (cf. section 4.5). Users of AnthroGraph can select one or several inventories among those specified in project management (cf. section 5.3). The choice is automatically limited to those inventories providing the required information. An investigation plan (cf. section 4.5) needs to be specified as a textual description and possibly as a digital asset. As an alternative or addition to the selection of inventories, users can select among the investigations specified in project management as sources of input.

Specimen collection takes the material designated by the skeletal inventories selected during the planning stage as input. If this material is selected or processed for investigation, the output is a specimen. If no specimen is defined, the material is rejected and the investigation aborted at this stage. Assays can be specified as part of the specimen collection process to support the decision whether material should be included or rejected. In both cases, specimen collection is to be documented in free text and possibly by means of digital assets (e. g. photographs). Alternatively, specimen and documentation of their inclusion or production can be selected from the investigation or investigations selected during the planning stage.

How essays are to be performed is documented in an investigation's study design. Screens for entering assay output need to provide these specifications which may contain text and possibly images (e. g. diagrams) as links, pop-up windows, mouse-over information or similar representations. Output can be in the form of numbers, categories or free text. Digital assets may accompany this information, for example in the form of photographs, x-ray images or machine output like spectrograms. The number of assays in an investigation may vary considerably, so input screens need to be designed in a way to accommodate large quantities of controls.

The data items generated by the execution of assays might be transformed for further analysis. Data transformation may take several forms, e. g. transformation from one scale to another or performance of calculations like multiplication with some factor or calculating the arithmetic mean of several values. Transformation processes may have several values as input and several values as output. Just like assays, they are documented with the study design and these specifications need to be accessible from the input screen.

Conclusions are inferences from data that require human reasoning. Their output may include numbers, categories or free text but should always be documented with a textual explanation.

Data transformations and conclusions pose a challenge to the development of a generic GUI as their numbers may vary and they might require each other's output as input. This might lead to complex workflows that are difficult to follow on a generic interface. Solutions to this problem might involve the display of processes for which a certain output can be an input or the definition of display sequences. Investigations with very complex workflows may require their own GUI templates. However, it is expected that in most cases data transformations will be performed before conclusions and that most investigations prescribe rather simple lines of reasoning.

Results from investigations (i. e. data and conclusions) may be picked up in written documents like reports or publications. To facilitate the understanding of investigations, AnthroGraph offers the possibility to specify which documents cover which elements. To this end, users need to be able to select or define documents and specify the items they are about.

5.5 Digital Assets Management

Digital assets may accompany several types of research data (cf. previous sections). AnthroGraph needs to provide the related documents in a file store for easy reference. Ideally, institutions deploying AnthroGraph should provide them under the URL that is used as an URI for these documents. More important, however, is their documentation as this information will stay with the research data when they are exported.

A generic screen for uploading files and annotating them with metadata needs to be complimented by a more specific screen for images in AnthroGraph MVP. Options for defining semantic relations with other items in the database need to be provided both in these screens

and in places where such connections are likely to occur.

5.6 Perspectives for Further Development

Development of AnthroGraph MVP requires to find simple navigation mechanisms that effectively support researchers in most of the tasks in osteological research. Their design needs to be simple and easy to understand in order to provide a basis for refinement in later versions. The scope for such improvements is quite unlimited. Specialised versions might support workflows typical to specific sub-disciplines of biological anthropology. Management of data from contract workers in bioarchaeology would require additional procedures for processing of contracts and quality management. Forensic use cases require integration with police investigations and court proceedings. More advanced deployments scenarios involving separated workspaces (section 2.3) will also involve modelling of additional operating sequences for their management.

6 Data Input, Accessibility and Output

In order to fulfil its purpose, AnthroGraph needs to cover three main application modes: data input, data search and data output. Data interaction must be open to users of varying computer literacy. In order to be accepted by all researchers, there have to be easy-to-use tools that let users quickly enter data and grab their datasets for processing elsewhere. At the same time, more intricate data interactions need to be made possible for application in professional data management.

6.1 Data Input

AnthroGraph provides a work environment for researchers to manage and document their osteological investigations without the need for learning about the technologies involved. This is important for a broad introduction of the software in biological anthropology and reaching out to researchers who are primarily concerned with anthropological analyses rather than data management. The GUI needs to support the workflows sketched out in chapter 5 at a usability that at least equals that of other applications like Osteoware and AnthroBook.

While these are merely interfaces for data tables, however, AnthroGraph has to meet additional challenges. It needs to automatically create and record the semantic coherence of the pieces of information entered through the controls. This also needs to process session and system variables like identity of the currently logged-in user, current date and time etc. User accounts need to have a respective RDF representation of the person holding the account which can be automatically identified as author of data input at a given moment. Another challenge is the representation of extensions whose exact content cannot be predicted. The GUI has to adapt to their specifications and provide the correct controls for their usage. Finally, AnthroGraph is to bring the novel functionality of managing digital assets, born of the ambition to represent these and their meaning for the documented investigations in the re-

search data output.

It should be noted that the AnthroGraph work environment mostly works as a specialised RDF data editor to create code that could also be produced otherwise. RDFBones data sets can be produced with text editors or more specific editors for semantic data (e. g. Protégé). Some use cases might profit from partly automated code generation. Whatever the method of its production, well-formed code can be imported directly into AnthroGraph, allowing for further editing through the GUI.

A prominent potential of RDF is the continuing harvest of data from other resources. This involves repeated queries with external databases and transformation of results into RDF statements. In this way, the information system can incorporate other established databases and provide their contents for reference. While such imports might be beneficial for many use cases, import filters need to be defined with the underlying database system and are outside the scope of AnthroGraph.

6.2 Data Search

AnthroGraph needs to provide functionalities to look up, filter and search data on the basis of collections, series of material, research projects, investigations, researchers, institutions etc. Overviews of tabular data need to be provided with investigations and projects for quick reference.

Search functionalities can be realised as keyword or faceted search. The software should also contain an interface for running custom SPARQL queries but it depends on a deployment's data access policy if this should be provided to all users or even made public (cf. chapter 7).

Display of data can follow different principles than data entry and should maximise clarity.

6.3 Data Output

Researchers will mostly request data in tabular form for further processing with other applications. Output for long-term storage in repositories should be in RDF format to preserve maximum information density. In this context, output as mixed data models, containing both ontology and instance information, is a desired feature.

6.3.1 Generic Output

Researchers should be provided with a quick way to access the data they have entered. Therefore, immediate download of data from individual inventories and investigations needs to be provided through the GUI. These datasets will be based on generic queries covering requirements of most common use cases. Extensions should provide such generic queries to provide an output that is meaningful in the context of the methods they implement. For the development of AnthroGraph, this implies that queries provided with ontology extensions need to be evaluated in the GUI. The application as such might provide generic queries fitting overall data structures from the RDFBones core ontology as fallback solutions.

6.3.2 Custom Output

Data output that exactly matches the requirements of a specific research objective need to be generated through SPARQL queries. Who can run these and in which circumstances is determined by a deployment's business model (cf. section 6.2).

6.4 Perspectives for Further Development

Improvements for data input that can be envisaged for the future include means of offline data entry, e. g. through PDF forms or templates for the data acquisition software OsteoSurvey for mobile devices.

Also, interfaces with other applications are likely to become requested. The statistical platform R is proven to interact well with open SPARQL endpoints. Pushing data to the application programming interface (API) of the collaboration research platform CoRA might open up provision of data to applications hooked up with this system. Perspectives in this direction are expected to unfold with the current activity related to data standardisation and management in biological anthropology.

7 Access Restrictions

Securing research data from unauthorised access is a prominent concern with researchers and institutions in biological anthropology. This creates the demand for fine-grained access control. AnthroGraph MVP, however, does not address this problem in depth and focuses on applicability in research instead. Access control is especially important for deployment with research institutions, collections and data platforms (cf. section 2.3). Exact requirements will be evaluated in phase 7 of the proposed work programme.

For AnthroGraph MVP, it suffices to realise access control among logged-in users on the level of page visibility. The underlying deployment scenario (research projects, cf. section 2.3.1) assumes that all users holding accounts for the information system are generally trustworthy collaborators acting with good intent and does not anticipate hacking attacks from logged-in users.

AnthroGraph requires that users are represented by an instance of type 'Person' in the information system's knowledge graph. This is necessary to automatically define this person as author of data entries (e. g. annotations, cf. section 4.7).

7.1 Public Data

Institutions deploying AnthroGraph might want to make selected data publicly accessible. Examples for such information include catalogues of their holdings, archival information governing the potential use of the material and information on types of investigations for which research data are available. Publishing such information helps researchers with planning their studies and leads to more precise requests and reduced administration efforts on the part of collection curators.

To realise this demand, AnthroGraph needs to provide the option to make selected pages displaying data publicly while keeping the rest exclusively accessible for logged-in users.

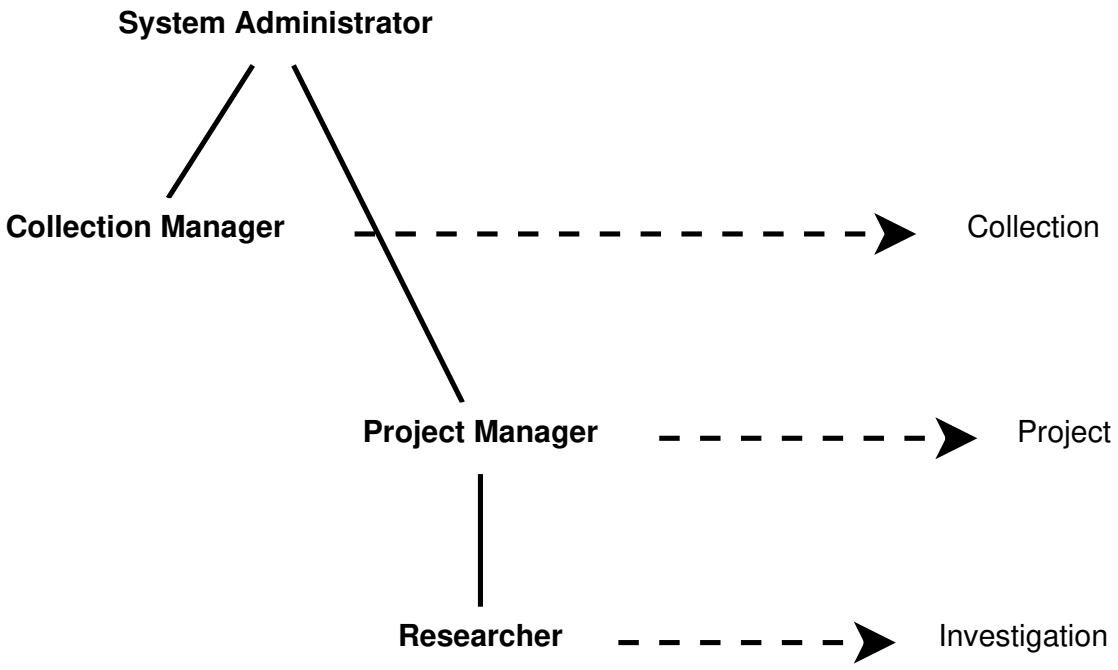

Figure 7.1: Hierarchy of roles (bold type) in AnthroGraph MVP and their primary concerns.

Public pages should also help institutions to publish general information like terms of usage or request forms.

7.2 Internal Access Control

Among logged-in users, AnthroGraph MVP needs to provide access control on the basis of roles (figure 7.1) in order to enable users to take responsibility for their data entries. If all data were editable by all users, authoring of data would be rendered meaningless. On the other hand, several persons should be able to correct obviously erroneous data entries, even when the original author is not available. To avoid confusion, it should be noted that roles can be assigned on two different levels:

1. RDFBones provides a role model classifying the contributions of individuals to research projects and the management of institutions. These refer to a real persons and are not restricted to their status as users of the information system.

2. Application software commonly defines roles to restrict rights and actions of individual users of the software. These refer exclusively to their status as application users and do not necessarily mirror their roles in real life.

Both role models can potentially be used to restrict rights of logged-in users to see certain application pages. Here, requirements for the use of roles in AnthroGraph MVP are related without implying how and on which level they are eventually implemented.

A particular responsibility is to keep information on the collection on which a certain deployment of AnthroGraph is based coherent and up to date. Therefore, collection management (cf. section 5.1) should be reserved to a group of users representing the deploying institution for them to be able to take responsibility of the material's curation and correct representation. This necessitates a separate workspace exclusively accessible to these collection managers.

Researchers conducting investigations should be able to take responsibility for their documentation. To avoid cross-editing of data within project groups, existing data entries should only be editable by their authors, i. e. by the users who originally made these entries. To broaden the spectrum of people who can edit research data, projects can have managers that have write access to all investigations within a project. They are also charged with providing general information on the project (cf. section 5.3). Project groups that prefer to collaborate on all investigations, can declare all their members as project managers. To facilitate work coordination, initiation of new projects and investigations should be reserved to the role level above, so that only system administrators can create new projects and only project managers new investigations. While doing this, they can specify who will be responsible for the newly initiated processes.

Implications of these restrictions are sketched out in figure 5.1.

7.3 Perspectives for Further Development

As mentioned above, access restrictions have been identified as a central issue for all target groups of AnthroGraph. It remains to be analysed what concrete requirements exist and to what degree AnthroGraph can and should support them. This will have to be done in close collaboration with potential adopters and specialists for the deployment of e-Research Technologies.

A future challenge might be the provision of workspaces to projects and collections that are to some degree guaranteed to protect the data produced therein. It needs to be determined which level of security is actually required and the possibility to analyse pooled data from different such workspaces must be ensured.

8 Extensions Management

A key feature of AnthroGraph is customisability by means of RDFBones ontology extensions (cf. section 3.1). In order to provide optimised interaction with concepts contained in ontology extensions, however, they might be just parts of application extensions with a larger scope. Application extension should have the following components, some of which might be optional.

- Version information for application extension
- RDFBones ontology extension
  - version information for ontology extension
  - SPARQL queries for generic data output
- Digital assets
  - sample data
  - illustrations as part of assay documentation
- Custom page templates for classes introduced by the ontology extension

Ontology extensions consist of additional RDF code defining subclasses of classes from the core ontology and their semantic relations. They should also contain SPARQL queries defining tabular data output for instances of these classes (cf. section 6.3.1).

The information contained in the ontology extension is complimented by other data that cannot all be expressed in RDF. Assays may require stereotypic illustrations as reference for categorising osteological observations. These graphics need to be loaded into the information system in order to be displayed with assay documentation. Their metadata and semantic relations can be defined already in the RDF code of the ontology extension. Another possible component are page templates configuring data entry forms specific to the extension. This

makes sense for institutions streamlining their workflows and for extensions with complex successions of data transformations and conclusions (cf. section 5.4).

The various components of application extensions need to be integrated into the information system and these processes need to be reversed if an extension is to be removed from the system. While this task calls for automation, it is assumed that with AnthroGraph MVP most of extensions management will be done manually by system administrators (cf. section 7.2).

8.1 RDF Import

The RDF code of the ontology extension needs to be imported into the information system's database. With a standard data import, however, core ontology information and extensions information become inseparable, making removal of extensions difficult, if not impossible. Therefore, the knowledge graphs of ontology extensions should be deposited as individual files and evaluated from there. After removing such a file, the database needs to be updated.

Upon removal of an extension, research data generated through this extension will still remain in the database. With the information from the extension missing, however, they are no longer fully documented. Removal of all related instance data could be effectuated by extensions providing SPARQL update queries to this purpose. The problem here is that extensions are encouraged to borrow each other's elements so that unwanted data deletion would be a threat.

To avoid deinstallation of extensions, AnthroGraph could offer a functionality to deactivate installed extensions to the effect that they are no longer offered to researchers for documentation of investigations.

8.2 Digital Assets Management

As all semantic information relating to digital assets is contained in the ontology extensions, the asset files just need to be saved in a place where they are recognised by the file store.

8.3 Page Template Management

How configuration files for pages are handled depends on the framework used for implementing AnthroGraph. There needs to be an option to add and remove templates provided with application extensions.

8.4 Perspectives for Further Development

With wide-spread adoption of AnthroGraph, it will be indispensable to develop functionalities for automated extensions management. This might involve introduction of a container format that helps to provide extensions as single files. These efforts, however, are only justified if intense usage of the software is secured.