Data Curation: How and Why. A Showcase with Re-use Scenario

PHILIPP GERTH, ANNE SIEVERLING AND MARTINA TROGNITZ
German Archaeological Institute, Berlin, Germany

The IANUS project, funded by the German Research Foundation (DFG), is building a digital archive and portal for archaeology and ancient studies in Germany. Following a 3-year phase of conceptual work, the archive and portal are now being implemented and the data center is beginning its operational work. Data curation is essential for preservation of digital data and helps to detect errors, aggregate documentation, and ensure the reusability of data; in some cases, it can also add useful additional files and functionality. This paper presents the workflow of data curation based on a data collection about European vertebrate fauna. It exemplifies the different stages of processing a dataset at IANUS according to the OAIS model – from its initial submission until its final presentation on the data portal. Data access and reusability can be enhanced by enrichment, in the case of the vertebrate fauna dataset, by GIS integration of geographic information and reutilization of bibliography. Furthermore, a data re-use scenario is presented in which the dataset has been integrated with one from another repository by using Semantic Web technologies.

Key words:
Data Curation, Data Enrichment, Long Term Preservation, Data Re-Use, Semantic Web.

SDH Reference:
Philipp Gerth et al. 2017. Data Curation: How and Why. A Showcase with Re-use Scenario. SDH, 1, 2, 182-193.

DOI:10.14434/sdh.v1i2.23235

1. INTRODUCTION

IANUS is a research data center for archaeology and ancient studies in Germany that provides a digital archive and a portal for data dissemination. The project is funded by the German Research Foundation (DFG) and is divided into two phases. The first was a three-year phase of conceptual work; in the current second part, the concepts are being implemented and the data center is beginning its operational work. The aim is to establish a reliable and sustainable data center where archaeologists as well as ancient studies researchers in general can archive their research data, and start new research projects with the data they will find in the IANUS data portal [Schäfer et al. 2015, pp. 131-134].

The work for this article has been carried out in the project IANUS, which is funded by the DFG under grant agreement no. 903456 and in the ARIADNE project, which is funded by the European Commission under the Community's Seventh Framework Programme, contract no. FP7-INFRA-2012-1-313193.
Author's address: Philipp Gerth, Anne Sieverling and Martina Trognitz, IT Department, German Archaeological Institute, Podbielskiallee 69-71, 14195 Berlin, Germany; email: (philipp.gerth, anne.sieverling, martina.trognitz)@dainst.de

Recently the data portal was launched to present the first datasets from the digital archive (available at http://www.ianus-fdz.de/datenportal/). The homepage of the portal gives an overview of the available datasets, showing a preview picture and a snippet of the project description. Each of the actual datasets has its own homepage, providing the project overview, and two subpages with rich metadata and the files available for download.

Figure 1. Screenshot of the data portal with the homepage of a dataset.

Information about the data provider(s), a map, the associated institution holding the copyright, the license conditions (e.g. CC-BY or CC-BY-SA), the digital object identifier DOI [Trognitz 2013] and a recommended citation can be found on the left sidebar of the dataset pages (Fig. 1). The homepage of a dataset displays the abstract of the project, the description of the dataset, selected publications, a statement of the data provider and relevant keywords. The subpage *metadata* presents the keywords concerning subject, content, localization, chronology and method, related publications, and statistical information about the files. Before viewing the *data* page, the terms of use have to be accepted. This page displays all available files of the project, which are organized in directories. By clicking on the folders on the left side in the box named *project structure*, a detailed overview of the files is shown and the files can be previewed or downloaded. Before the download starts, a window with the license conditions informs the user about the terms of re-use. In some datasets, an additional ZIP file is provided to allow downloading the whole dataset at once.

Before a dataset can be presented on the portal data curation is needed. The curation process begins with a systematic review of the dataset. The review provides the basis for converting files into appropriate formats for long-term preservation, preparing documentation and metadata, detecting possible errors, and other measures to ensure the preservation and reusability of digital data.

In some cases, further functionality and additional files can be created to enrich the dataset during the process of data curation. We present the workflow of data curation based on a data collection about European vertebrate fauna (available at http://dx.doi.org/10.13149/001.mcus7z-2), as well as its enrichment, and showcase a possible re-use scenario of this dataset.

2. PROJECT AND DATASET DESCRIPTION

The dataset "European vertebrate fauna" was produced in a project in the 1990s by three scientists at different German institutions: Norbert Benecke (German Archaeological Institute), Angela von den Driesch (University Munich) and Dirk Heinrich (University of Kiel). In the project, information about European vertebrate fauna was collected and its development from the late Pleistocene until the Middle Ages analyzed, covering a time span of more than 10,000 years. No new animal bones were gathered and analyzed, but the data about all already published animal remains across Europe were entered into a dBase-database. In 4500 publications, over 8200 find spots and 100 different species were documented. On the basis of this huge compilation of data the researchers investigated changes in skeletons, habitat, human-animal relations and many other aspects. See for example [Crees et al. 2016; Sommer et al. 2014] among other project publications.

All parts of the project, including the produced data, were excellently documented and even the structure of the database was published [Benecke 1999, pp. 152-154]. But subsequently, after later examination of the data by different scientists, the database was not used anymore. Instead, the data was exported into tables, which were used and subsequently changed. This is the reason why the database does not represent the actual state of progress any longer. Therefore, IANUS did not receive a database, but a dataset containing the different exported and updated tables and files. These files comprise the largest part of the dataset, and their organization is based on the original structure of the database, consisting of the topics catalogue, species, measurements and literature. The structure of the data sheets, their connection and content as well as the abbreviations, are explained in a

detailed readme file that helps in understanding the dataset and curating it in a proper way [Benecke et al. 2016].

3. DATA CURATION

The curatorial work was carried out and documented according to the *Open Archival Information System - OAIS* https://public.ccsds.org/pubs/650x0m2.pdf [CCSDS 2012]. The *Submission Information Package* (SIP) of the data provider was transformed into a valid *Archival Information Package* (AIP) and for the data portal into a *Dissemination Information Package* (DIP) (Fig. 2).

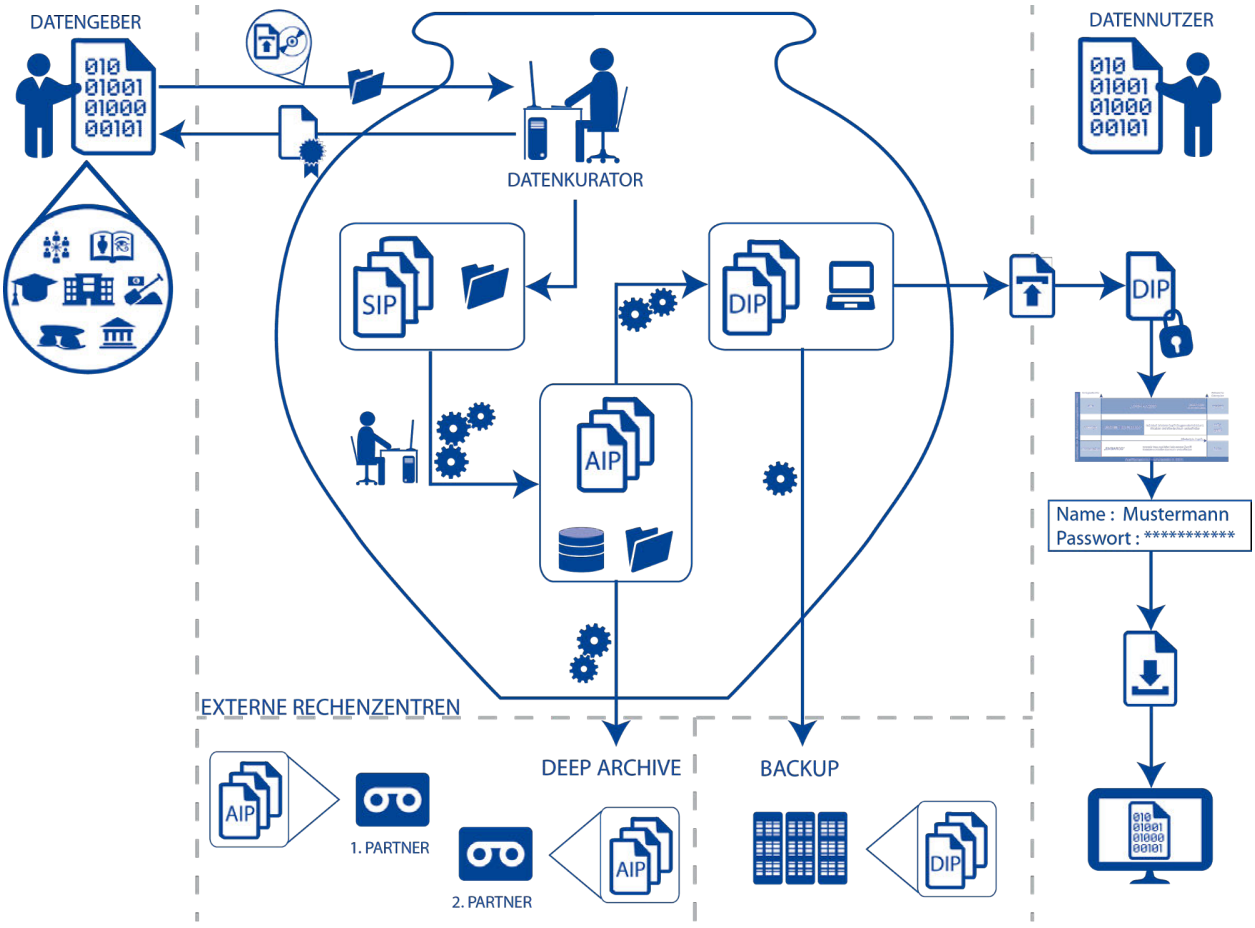

Figure. 2. Workflow of data submission, curation and dissemination according to the OAIS Model.

3.1 File renaming and conversions

To generate the AIP and DIP a curation strategy based on IANUS' IT-Recommendations [IANUS 2016] was defined and the following file conversions were executed:

Some of the files and folders had umlauts and space characters that were changed with *Bulk Rename Utility* (available at http://www.bulkrenameutility.co.uk/Main_Intro.php).

For the AIP, doc-files were converted to DOCX. This was done with the tool *doc2docx* (available at http://www.er-ef.net/doc2docx.html). For digital preservation and especially for dissemination, docx-files were also saved as PDF/A-1, using *Adobe Acrobat X*. Format validation of the PDF/A-1 files was additionally executed with *veraPDF* (available at http://verapdf.org/).

The tables had to be converted from XLS to XLSX for the AIP as well as for the DIP. Additionally, a conversion from XLSX to CSV was carried out. For the batch-processing from XLSX to CSV *bytescout spreadsheet* was used (available at https://bytescout.com/).

All xml-based files were validated with the *Open XML SDK 2.0 Productivity Tool* of Microsoft (available at https://www.microsoft.com/en-us/download/details.aspx?id=5124). For future validation, it is planned to test other, non-proprietary tools and integrate them into the data curation workflow.

3.2 File reorganization

The main data provider, Prof. Dr. Norbert Benecke, allowed us to change the structure of the folders and files as well as their names; therefore, we could restructure the file tree. The folder containing all readme files was deleted and the files were moved to the respective other folders to keep them with the contents they describe (Fig. 3). In the SIP, the folder "publications" contained a single file in DOCX format with a list of 13 publications that were published during the lifespan of the European vertebrate fauna project. This list was included in the project metadata and the file was converted to PDF/A. For the resulting AIP and DIP the folder was deleted and the file moved to the uppermost level of the whole dataset [Benecke et al. 2016].

Since the dataset does not contain a database, as explained above, the names of the folders were changed manually from 'database' to 'files' so as to not confuse future users. For example, the folder 'database countries' was renamed to 'catalogue files' and 'database species' to 'fauna files' (Fig. 3).

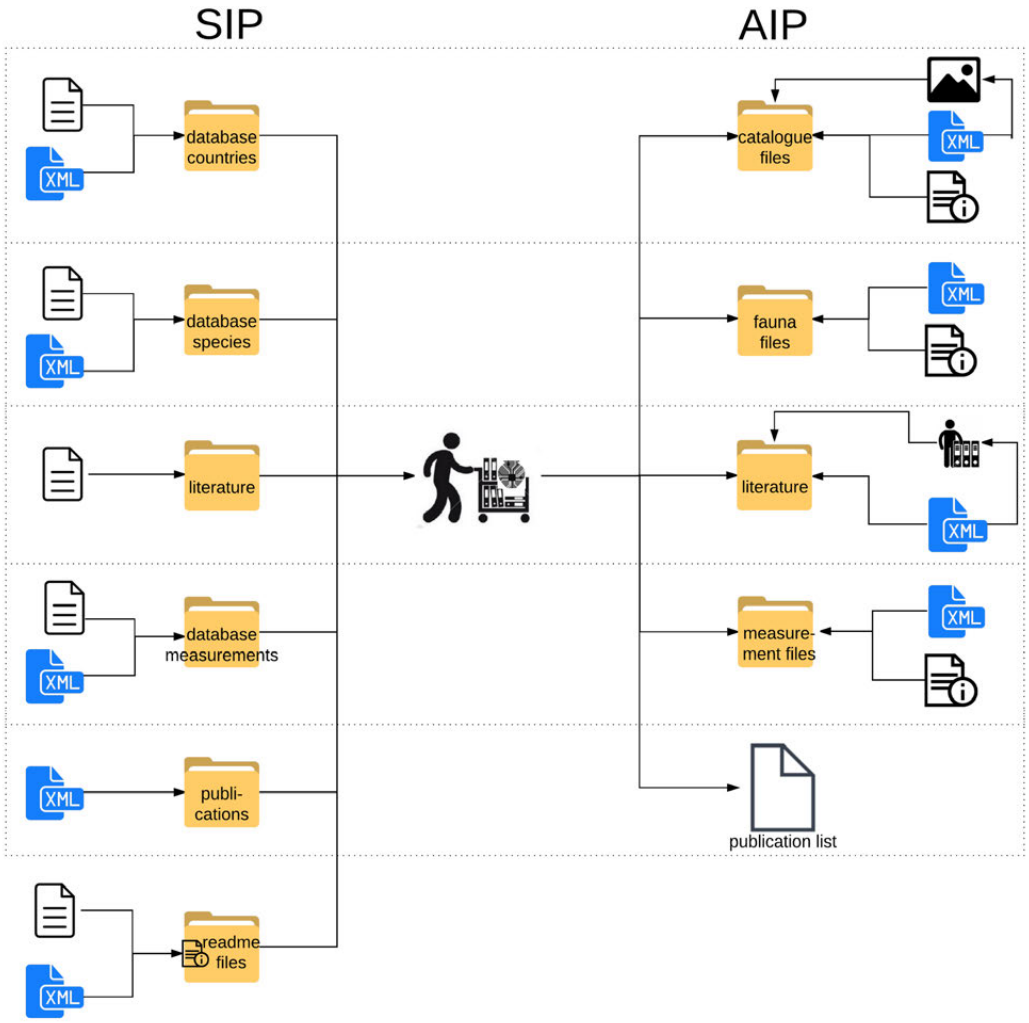

Figure 3. Changes and enrichment of the dataset from SIP to AIP.

4. ENRICHMENT OF THE DATA

While restructuring and converting the dataset, ideas came up to enrich the dataset in order to facilitate re-use. Two additional means for an easier access were generated: one for GIS and one for reference management.

4.1 GIS integration

To prepare GIS integration, a single file containing all geographic information from the 35 files of the folder "catalogue files" was created [Benecke et al. 2016]. This file was then imported into the open source Desktop Geographic Information System *Q-GIS* (available at http://www.qgis.org/de/site/).

Import and display of the find-spots data revealed that the coordinates in the tables needed revision. The coordinates were expressed in degrees, minutes, and seconds (DMS) but used the notation of decimal degrees (e.g. "15,59" instead of 15°59'). Therefore, the coordinates were converted [Linoff 2015, 148] into proper decimal degrees (e.g. 15°59 to 15,9833) for use in Q-GIS (Fig. 4). Errors in some coordinates were corrected manually and documented in a supplementary readme file. All resulting files (CPG, DBF, GEOJSON, PRJ, QPI, SHP, SHX) generated from Q-GIS were also stored and are provided together with the original data for re-use [Benecke et al. 2016].

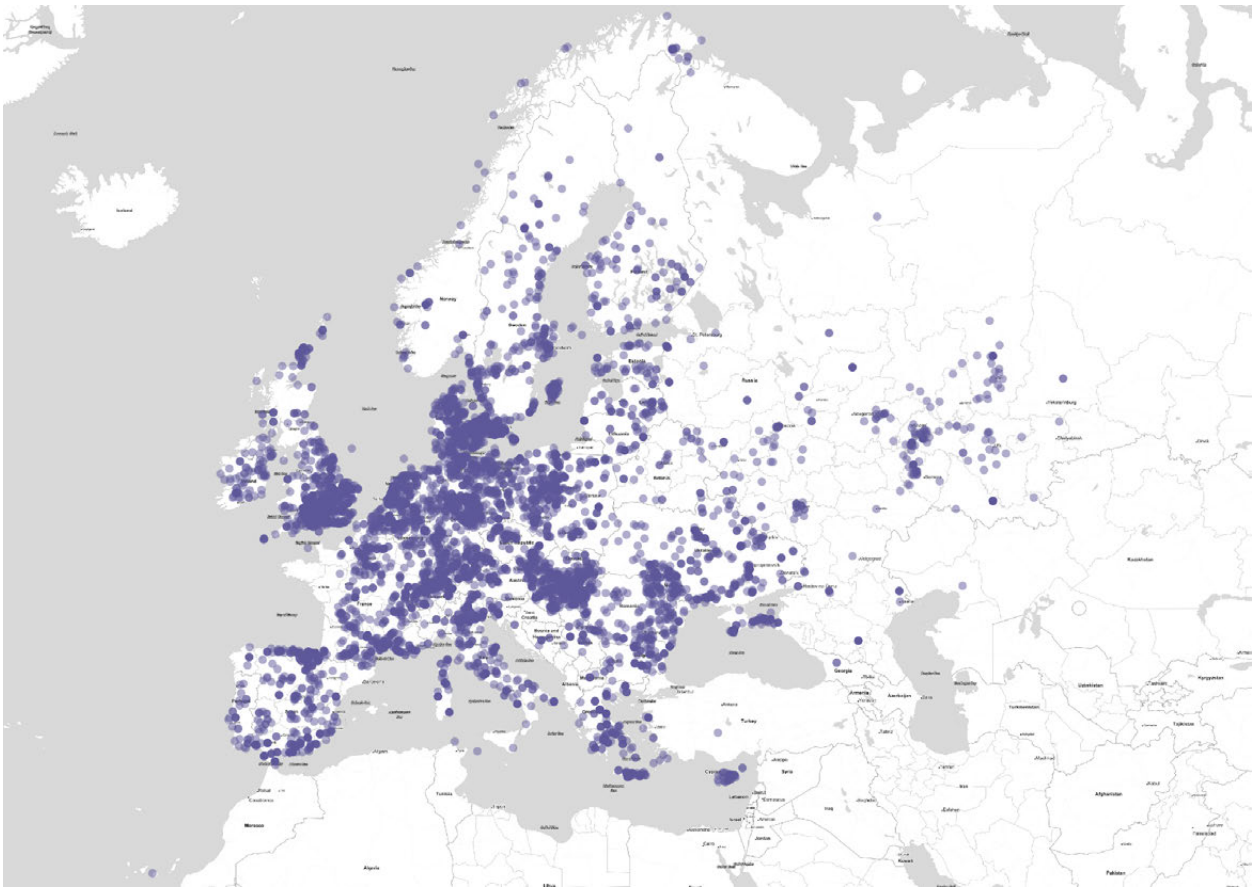

Figure. 4. Distribution map of the European vertebrate fauna find-spots.

4.2 Bibliographic information

In addition to converting the original doc-files containing bibliographic references from the folder "Literature" into its respective archival formats (DOCX and PDF/A), it was decided to add further functionality by aggregating the references dispersed across 22 files into one single bibliographic file in BibTeX format. This allows for integration of the information into reference management software.

For this purpose, the documents were saved manually as plain text encoded in UTF-8. With a Python script, the text files were scanned for information about author, year, and title, and converted into a bib-file. The original reference was kept and added into the note field. During this process missing information, wrong punctuation and similar errors were manually corrected in the text files. In the resulting bib-file parts where data providers had marked missing information with question marks or 'xxx' were revised and completed when possible; duplicates were also removed. A single file containing the whole bibliography of the dataset in a standardized format increases the re-use potential of this information [Benecke et al. 2016].

5. DATA RE-USE ON THE EUROPEAN LEVEL

In order to demonstrate the potential of integrating different scientific datasets in the domain of archaeological science, two heterogeneous zooarchaeology datasets, one hosted by IANUS and one by the Archaeological Data Service in York/UK, were combined by using Semantic Web technologies. Researchers of zooarchaeology have a long tradition of sharing their datasets and articles in community portals like *Bone Commons* http://alexandriaarchive.org/bonecommons/, the Zooarchaeology social network, the *ZOOARCH* email discussion list https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=ZOOARCH and other platforms [Kansa and Deblauw 2011]. This discipline serves as a useful case study, as the terminology used is highly standardized, its materials and methodologies are global in scope and many research questions are only answerable by taking into account multiple datasets. The integration of the two datasets was undertaken as part of the EU project ARIADNE, which brings together and integrates existing archaeological data infrastructures with the goal of offering researchers unified search and discovery facilities over a wide range of distributed datasets [ARIADNE 2016].

Besides the already described European vertebrate fauna, a second dataset was chosen from the project "A Review of Animal Bone Evidence from Southern England," funded by English Heritage and aimed at reviewing animal bone evidence from the Late Bronze Age throughout Late Iron Age in southern England. The Regional Review report [Hambleton 2008], for which this database serves as an open accessible online appendix, provides a synthetic review of published faunal assemblages. Consequently, analyses (e.g. ageing, butchery, biometric data) focus on the exploitation and deposition of sheep, cattle, pig, horse and dog. Other taxa (e.g. wild mammals, birds, fish and amphibians) are also discussed. The information in the database, published by the Archaeological Data Service [Hambleton 2009], is based on 108 site reports, which correspond to excavations at 101 separate monument locations and 154 distinct 'assemblage' records for faunal assemblages. Additionally, bibliographic references for all zooarchaeological reports reviewed are listed in the database.

The two datasets were mapped to the common super-classes and relationships of the ontology CIDOC-CRM to relate both relational data models to a common standard. To guide these mappings, the tool *3M Mapping Memory Manager* (available at www.ics.forth.gr/isl/3M/) was used. The knowledge graph derived from the mapping and alignment of the two datasets is depicted in Fig. 5. To overcome the language barrier between the German and English datasets a common standard was introduced, by using the *Encyclopedia of Life* (EOL, available at http://eol.org/), which provides

biological definitions of species in a classification tree. As the thesaurus is not available in Semantic Web formats, the terms used were described with RDF/XML.

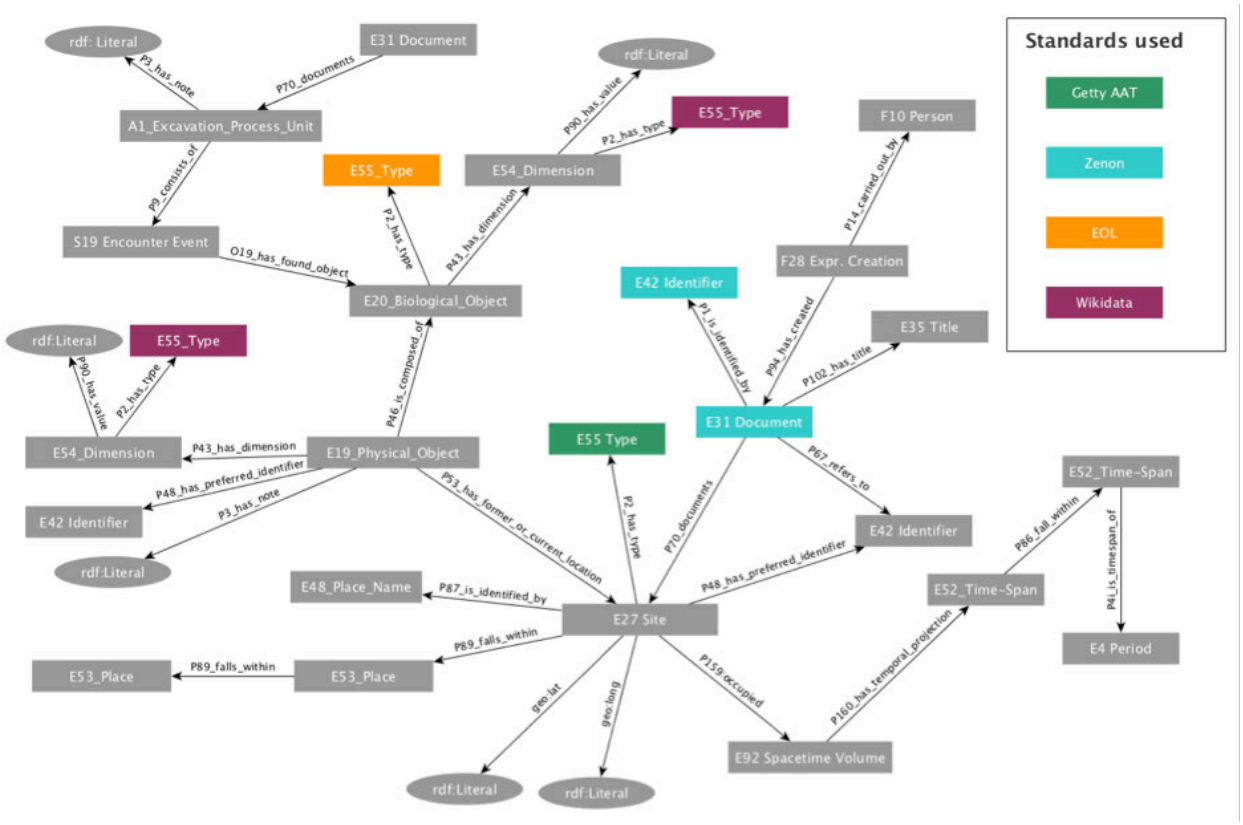

Figure 5. Graphic representation of the CIDOC-CRM mapped datasets in the "Animal Remains" scenario

The integration of the two zooarchaeology datasets leads to an environment where users are able to specify queries that run on a common aggregated repository and combine the results coming from the different datasets (Fig. 6). The first example in the figure is a species-centric query, where all sites with horse remains are shown with their bone assemblages. For a researcher who is interested in the distribution of a specific species the literature references could be a welcome starting point. The second example illustrates possible basic statistical analyses with the help of the query language SPARQL over a common mapped dataset. A frequency distribution of selected species in the archaeological contexts is depicted in the bar chart. This approach highlights the high potential of re-using datasets and how new results can be gained without having to invest in the costly acquisition of new research data.

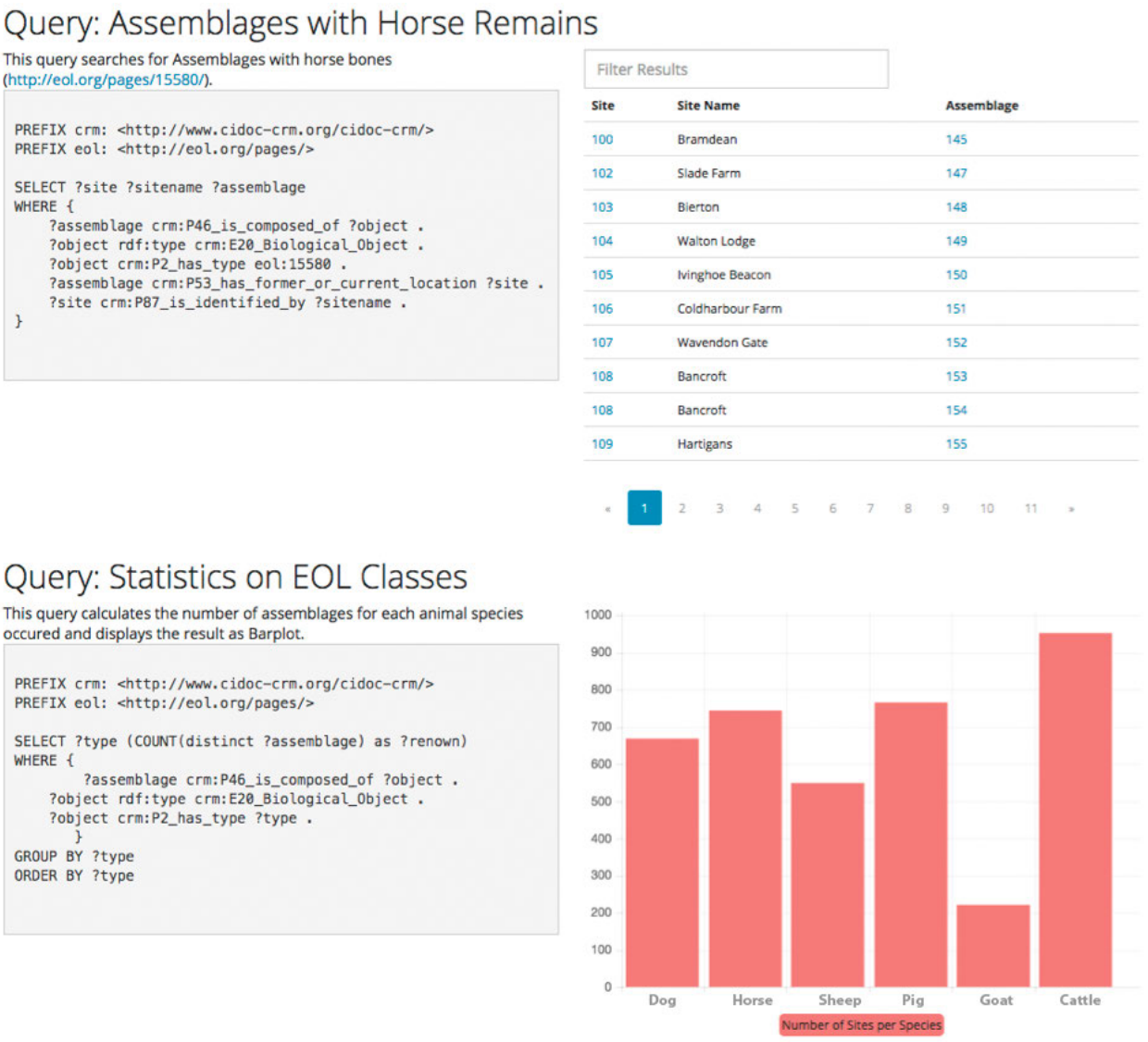

Figure 6. Queries on the two integrated zooarchaeology datasets: a SPARQL query on the left side and visualization of the result on the right side

6. DISCUSSION

This paper presented the data portal of the IANUS project and addressed some details of the data curation workflow according to the OAIS standard. A zooarchaeological dataset was used to exemplify the steps involved in the creation of a dataset suited for long-term preservation.

Two cases of enriching the data were presented, to allow for GIS and reference management integration. During this process, some errors in the dataset were detected, corrected and documented. This led to a significant improvement of data reusability without violating the

archiving principles of the immutability and authenticity of data, as new datasets were created while still preserving the original ones.

Finally, we have shown a solution for the integration of heterogeneous datasets using Semantic Web technologies. These aggregation activities point to a very promising direction and could incorporate all archaeological datasets stored in different data centers into an integrated knowledge graph to provide access to a huge amount of comparable data. This could enable users to answer research questions across heterogeneous resources through gaining statistically more reliable results without the acquisition of new data. But to use this potential, willingness of the researchers to make their data openly available in open and standardized formats is necessary.

7. REFERENCES

ARIADNE – Advanced Research Infrastructure for Archaeological Dataset Networking in Europe. 2016. Building a Research Infrastructure for Digital Archaeology in Europe. ARIADNE Booklet, December 2016. Retrieved March 18, 2017 from http://www.ariadne-infrastructure.eu/About

N. Benecke. 1999. The Project "The Holocene History of the European Vertebrate Fauna." In N. Benecke, ed. 1999. The Holocene History of the European Vertebrate Fauna. Archäologie in Eurasien 6, Rahden/Westfalen: Marie Leidorf Verlag, 151-161.

N. Benecke et al. 2016. Holozängeschichte der Tierwelt Europas [data-set], Berlin: IANUS. DOI:http://dx.doi.org/10.13149/001.mcus7z-2.

Bone Commons, http://alexandriaarchive.org/bonecommons/

CCSDS - Consultative Committee for Space Data Systems. 2012. Reference Model for an Open Archival Information System (OAIS). Recommended Practice CCSDS 650.0-M-2. Magenta Book, Washington, DC: CCSDS Secretariat. Retrieved March 18, 2017 from https://public.ccsds.org/pubs/650x0m2.pdf

J. J. Crees et al. 2016. Millennial-scale Faunal Record Reveals Differential Resilience of European Large Mammals to Human Impacts across the Holocene. Proceedings of the Royal Society B, 283: 20152152. DOI:http://dx.doi.org/10.1098/rspb.2015.2152.

E. Hambleton. 2008. Review of Middle Bronze Age - Late Iron Age Faunal Assemblages from Southern Britain. Research Department Report Series number 71-2008. English Heritage.

E. Hambleton. 2009. A Review of Animal Bone Evidence from Southern England [data-set]. York: Archaeology Data Service [distributor]. DOI:http://dx.doi.org/10.5284/1000102.

IANUS, ed. 2016. IT-Empfehlungen für den nachhaltigen Umgang mit digitalen Daten in den Altertumswissenschaften [Version 1.0]. DOI:http://dx.doi.org/10.13149/000.y47clt-t.

S. W. Kansa and F. Deblauwe. 2011. User-Generated Content in Zooarchaeology - Exploring the "Middle Space" of Scholarly Communication. In E. Kansa et al., eds.. Archaeology 2.0 - New Approaches to Communication and Collaboration. Los Angeles: Cotson Institute of Archaeology Press, 185-206.

G. Linoff. 2015. Data Analysis Using SQL and Excel. New York: Wiley.

F. Schäfer et al. 2015. Forschungsrohdaten für die Altertumswissenschaften – eine kurze Bilanz der aktuellen Situation von Open Data in Deutschland. Archäologische Informationen, 38 (2015), 125-136. DOI:http://dx.doi.org/10.11588/ai.2015.1.26156.

R. S. Sommer et al. 2014. Range Dynamics of the Reindeer in Europe during the Last 25,000 years.

Journal of Biogeography 41 (2014), 298-306. DOI:http://dx.doi.org/doi:10.1111/jbi.12193.
M. Trognitz. 2013. Abschlussbericht Testbed "Persistent Identifiers." Retrieved January 20, 2017 from http://www.ianus-fdz.de/projects/ergebnisse/wiki
ZOOARCH, email discussion list, https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=ZOOARCH

Received March 2017; revised July 2017; accepted August 2017.