IJDC | *Peer-Reviewed Paper*

Publishing and Pushing: Mixing Models for Communicating Research Data in Archaeology

Eric C. Kansa
Open Context and University of California, Berkeley

Sarah Whitcher Kansa
Open Context

Benjamin Arbuckle
The University of North Carolina, Chapel Hill

Abstract

We present a case study of data integration and reuse involving 12 researchers who published datasets in Open Context, an online data publishing platform, as part of collaborative archaeological research on early domesticated animals in Anatolia. Our discussion reports on how different editorial and collaborative review processes improved data documentation and quality, and created ontology annotations needed for comparative analyses by domain specialists. To prepare data for shared analysis, this project adapted editor-supervised review and revision processes familiar to conventional publishing, as well as more novel models of revision adapted from open source software development of public version control. Preparing the datasets for publication and analysis required significant investment of effort and expertise, including archaeological domain knowledge and familiarity with key ontologies. To organize this work effectively, we emphasized these different models of collaboration at various stages of this data publication and analysis project. Collaboration first centered on data editors working with data contributors, then widened to include other researchers who provided additional peer-review feedback, and finally the widest research community, whose collaboration is facilitated by GitHub's version control system. We demonstrate that the "publish" and "push" models of data dissemination need not be mutually exclusive; on the contrary, they can play complementary roles in sharing high quality data in support of research. This work highlights the value of combining multiple models in different stages of data dissemination.

Received 27 October 2013 | *Accepted* 26 February 2014

Correspondence should be addressed to Eric C. Kansa, 125 El Verano Way, San Francisco, CA 94127. Email: ekansa@berkeley.edu

An earlier version of this paper was presented at the 9th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

 doi:10.2218/ijdc.v9i1.301

Introduction

Researchers and policy makers increasingly regard data management as a critical need in many areas of science, the social sciences and the humanities. Field research, particularly in archaeology, often spans these domains and typically generates large and complex databases, often describing historically unique topics. This makes data preservation a critical need. Though data archiving is of critical importance, data management needs extend well beyond "preservation for the sake of preservation." Editorial processes similar to conventional publishing can improve the quality and intelligibility of data and metadata, making datasets more easily understood and more comparable to other datasets. At the same time, many problems and issues in data only become apparent with reuse, especially analysis and comparison with other datasets. Thus, unlike conventional publishing's emphasis on producing fixed final products, data publishing can benefit from continual and incremental improvements supported by version control systems. To better understand these different data management needs, this paper discusses data publishing practices that supported a collaborative study involving data sharing, integration and analysis in archaeology.

Background

Policy changes, including Data Management Plans now required by the National Science Foundation (NSF) and certain programs of the National Endowment for the Humanities (NEH), highlight growing attention paid to research data. The White House Office of Science and Technology policy's early 2013 call for open access, as well as tentative steps toward greater openness in research data, further advance these policies.

While data is assuming greater importance in scientific policy, the research community still lacks consensus as to how to situate data management in scholarly communications and professional advancement. The NSF and NEH currently make no specific requirement for the management of data, leaving data management review criteria up to the discretion of review panels, which are mainly staffed by domain researchers. These reviewers often lack guidance or expertise in what constitutes a good data management plan. To help fill this void, several university libraries and disciplinary repositories have come together to give the research community better guidance in grant-mandated data management, such as the DMPTool[1], an online system to aid the creation of project-specific data management plans.

Though the DMPTool and similar services may help to improve practice, a general assumption remains that structured data mainly need to be "archived" with institutional or disciplinary repositories. In other words, a researcher's primary responsibility toward data currently centers on *preservation*. This emphasis on data preservation with institutional repositories represents a new normative best practice. In many ways, the idea that "data are for preservation" reflects an incremental change in the conduct of research. In this perspective, conventional refereed journal papers remain the primary vehicle of research communications, and data are mainly made available as supplements to support claims made in a paper. In principle, follow up studies can reuse archived data, but actual reuse of data remains rare (Wallis et al., 2013). Furthermore, in order to

1 DMPTool: https://dmp.cdlib.org/

encourage uptake among researchers, data repositories typically have very low barriers to accessioning datasets. This dual emphasis on preservation and ease-of-deposit means that scholars may archive datasets, but often those datasets have minimal documentation or processing to facilitate reuse.

While we agree with the necessity of archiving data with repositories, we question if such practices are sufficient to deal with the realities and complexities of data reuse. Archaeology is representative of the "small sciences" (Onsrud & Campbell, 2007), where research is typically conducted by single investigators or small teams, often in conjunction with regulatory compliance (government enforced mandates for environmental or cultural heritage protection). Methods and recording practices can vary widely, and are often tailored to meet the needs of different circumstances, including research agendas, and budget and time constraints (Dibble & McPherron, 1988; Eiteljorg, 1998). Such factors complicate data reuse. Thus, the small sciences (and digital humanities) should invest more thought and effort in proper data contextualization than that implied by simple data archiving approaches.

Methods

Because data archiving practices have only recently gained momentum in archaeology, the discipline still lacks a clear understanding of the factors governing data reuse. To explore the challenges of data reuse, we secured a grant from the Encyclopedia of Life[2] to bring a group of scholars together to publish and integrate data from 12 archaeological sites to explore research topics related to the origin and spread of domestic animals in Anatolia. This group, the Central and Western Anatolian Neolithic Working Group, represents a rare collaborative effort to publish and integrate open data in archaeology. Participants published faunal (animal bone) datasets from archaeological sites spanning the Epipaleolithic through the Chalcolithic (a range of 10,000+ years) in Open Context. Participants then analyzed subsets of the integrated data and presented their results to the group. Feedback from the data editing and integration process, and group discussions about using data produced by others, informed the results presented in this paper. While this paper focuses on the data management implications of this study, the project has also shed light on the development of Neolithic societies and the processes that initially brought agriculture to Europe (see Arbuckle et al., 2014).

Data Publication and Analysis Workflow

The Open Context editorial team (S. Kansa and E. Kansa) managed the process of data submission, editing and integration. The project director (B. Arbuckle) managed analysis of the integrated datasets. The following steps describe the process, which took place over six months from October 2012 to April 2013. In addition, email communications between Open Context and the data contributors were recorded and are being analyzed by the DIPIR project[3] in a study of data reuse. The data review, editing, annotation, analyses and publication steps below describe the project workflow (see Figure 1).

2 Encyclopedia of Life Computable Data Challenge grant: http://eol.org/info/345

3 Dissemination Information Packages for Information Reuse (DIPIR) project: http://dipir.org/. The DIPIR project explores scientific data reuse through systematic qualitative and quantitative studies of researcher interactions with data repositories.

 doi:10.2218/ijdc.v9i1.301

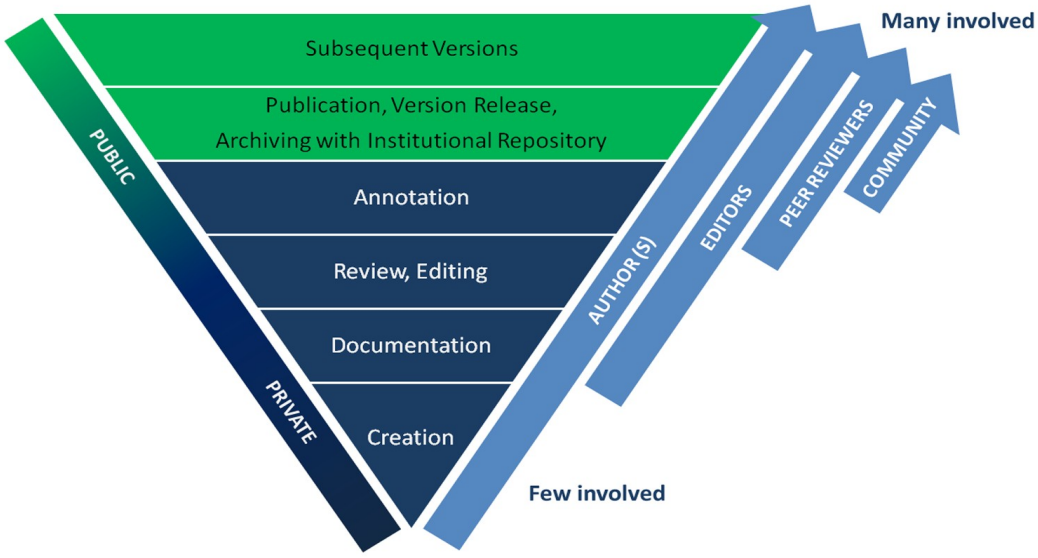

Figure 1. Data publication workflow, showing movement of content from private to public spheres and the key players involved in each stage.

Step 1: Solicitation and metadata documentation

After two years of informal discussion with colleagues, Project Director Benjamin Arbuckle invited zooarchaeologists[4] working in Turkey to share and analyze multiple datasets in a collaborative research project investigating the origin and spread of domestic animals in Anatolia. To maximize analytic freedom, the project leads requested full datasets rather than summarized data. We chose Microsoft Excel as a file format for submission because of its widespread use among the contributing researchers[5].

In order to facilitate citation as well as search, browse and retrieval features on Open Context, project leads requested specific accompanying metadata necessary for the datasets' reuse. This documentation included authorship information, basic project and site descriptions, keywords, relevant chronological ranges, and geospatial information for basic mapping. We also asked contributing researchers to include information on data creation methods and sampling protocols, and to describe each field of their submitted dataset. In most cases, contributing researchers submitted minimal documentation. Data editors needed to create supplemental documentation on behalf of contributing researchers, who then approved the additional information.

Step 2: Review, decoding and editing

Upon receiving a dataset (typically expressed as one or more spreadsheets), we began an initial stage of review, involving checks for internal consistency of datasets, especially for identifiers. In two cases, we found that submitted datasets contained non-unique primary identifiers. We resolved these issues with data contributors: in one case a researcher accidentally duplicated several records; and in another case, a researcher

4 Zooarchaeologists are individuals specializing in the study of animal remains from archaeological sites.

5 Because the datasets described zooarchaeological data collected in Turkey, Microsoft Excel was also chosen to avoid technical complications of character encoding. To promote interoperability and longevity, Open Context makes data available in UTF-8 encoded open formats (chiefly XML, XML-RDF, and CSV).

accidentally assigned duplicate identifiers to different bone specimens. In a third case, the initial data review stage led us to choose to delay publication of a dataset and not include it for comparative analysis in this study (see Discussion section).

Contributing researchers sometimes used coding systems as shorthand to facilitate data entry. For example, instead of typing "*Capra hircus*", an analyst may enter a coded value such as "7". There is no standard coding system used by zooarchaeologists, so each individual had to submit a code book for the translation of their data. That translation had to be done by hand because most of the researchers used Microsoft Excel and not a relational database, so coded values could not be automatically decoded through use of related "lookup" tables. In one case, we had to consult a 90-page code book (in PDF) to decode the values in a submitted dataset. Occasionally, we encountered undocumented codes and needed to consult data contributors. Data entry errors accounted for some problems, while others resulted from codes that simply lacked documentation, the latter requiring consultation with data contributors to explain their meaning.

The Open Context editorial team used Open Refine to perform basic checks and edits for each submitted dataset. These activities typically involved fixing spelling or capitalization inconsistencies, especially in classification fields. In addition, we checked numeric fields to see if they contained numeric values, and if not, we used Open Refine to correct and document such (non-numeric) values.

Step 3: Linked data annotation

Over the course of the project, we received data from 12 archaeological sites. Each dataset had its own unique organization (schema) and described zooarchaeological data using somewhat different terminologies and vocabularies. In order to make these datasets comparable, we annotated them with common ontologies. Ontology alignment included enabling cross-dataset comparisons with respect to taxa by annotating dataset-specific taxonomic categories with Web URIs for biological taxonomic concepts curated by the EOL[6]; annotating dataset-specific classifications of bone elements with URI-identified concepts curated by UBERON[7]; and using a controlled vocabulary developed by Open Context for bone fusion, sex determinations and standard measurements. None of the participating researchers had any prior familiarity with these ontologies. Annotation to these controlled vocabularies and ontologies provided the basis for data integration across the contributed datasets. Open Context's data editors (including a specialist in zooarchaeology) initially made the annotations while contributing researchers approved them. Using the conventions of "linked open data", web URIs identify concepts in referenced ontologies, and Open Context publishes these data in a variety of representations, including RDF (Kansa, 2012).

Step 4: Contributor review, peer review, and analysis

In collaboration with the authors of the datasets, Open Context's editors spent four months decoding and editing over 294,000 records of bone specimens from the 12 participating archaeological sites, and aligning the data to common ontologies. Upon review, contributing researchers faced little difficulty in checking the annotations against controlled vocabularies (especially EOL and UBERON) since they were

6 Encyclopedia of Life (EOL): http://eol.org

7 Uber Anatomy Ontology (UBERON): http://bioportal.bioontology.org/ontologies/1404. Though the project used UBERON initially for vocabulary control across datasets, we anticipate more sophisticated research possibilities that make use of semantic inferences based on the UBERON ontology.

 doi:10.2218/ijdc.v9i1.301

represented clearly in additional fields that could be sorted and filtered using tools like Microsoft Excel. Moreover, because each contributing researcher also participated in the analysis of edited and annotated data, they had ample opportunity to note problems in any of the 12 datasets. Each participating researcher then addressed a specific research topic using a subset of the data. Participants met in April 2013 at the International Open Workshop at Kiel University to present their analytic results on the integrated data and to prepare a multi-authored synthetic research paper (Arbuckle et al., 2014).

Step 5: Publication, indexing and archiving

After data contributors reviewed and accepted edits, annotation and metadata documentation, they communicated any change requests to the Open Context editors. We then published the edited and annotated data online, assigned persistent identifiers (DOIs), and entered the datasets also into a public version control system (GitHub), where all subsequent changes are publicly tracked and logged. Once published and indexed with Open Context, GitHub tracked further changes requested by participating researchers as well as outside researchers.

Upon publication, Open Context builds an elaborate index of all the data, metadata and annotations to facilitate a variety of faceted search and visualization functions and support a powerful application program interface (API). Open Context's APIs also make the data available to the California Digital Library for long term archiving. Open Context publishes the data as freely accessible, open data in a variety of formats, including XHTML (for viewing in web browsers), XML/RDF (for Linked Data applications), XML (for software parsing and GitHub version tracking), JSON (for visualization), and CSV (for convenient download and use in tools like Microsoft Excel).

Results

The workflows and methods described above transform "raw data" contributed by participating researchers into edited and annotated products ready for analysis. This workflow has informed some of the broader challenges in scientific data management and reuse. These are summarized in Table 1. Harley et al. (2010) noted widespread reluctance to share data, especially in archaeology. This project faced less reluctance, probably because Arbuckle had longstanding collaborative ties with the participants. Twelve of fourteen invited researchers agreed to participate. The two that declined saw their projects as still too "new" to share data and preferred to wait until (conventional) publication. Of the 12 researchers who participated, response times between data solicitation and submission varied widely, depending on the amount of clean up deemed necessary. No dataset was immediately ready for publication; that is, all participants needed time to prepare their datasets for dissemination (ranging from a few days to a few months). Thus while most researchers agreed to share data, most needed time and outside editorial assistance to assist in preparing data for reuse.

Table 1. Issues encountered in data publishing.

Stage	Issue	Resolution	Author Input Required?
Solicitation	Project too "new" to share publicly	Make agreement to publish data when "ready"; submit "forthcoming" project metadata	Yes
Metadata documentation	Incomplete metadata	Request from author	Yes
	Crediting data creators in large team projects	Create semi-automated means of assigning authorship order	Yes
Review, decoding, and editing	Non-unique primary identifiers	Custom scripting, Open Refine	Frequently
	Coded data	Decode data based on code book/sheet provided by author; custom scripting; Open Refine	Frequently
	Data consistency	Use Open Refine to clean data	Infrequently
Linked data annotation	Data annotation	Use of domain-specialist editor	Infrequently
Reuse/analysis	Insufficient information for analysis	Improve project metadata	Yes
	Poor data modeling practices	Improve recording and modeling practices to facilitate comparability of datasets	Requires change in data creation practices and management tools

Data Editing

We encountered a variety of challenges in editing contributed datasets. In two cases, submitted datasets were not detailed enough to include in the data integration phase of the project (for example, data tables containing summary data rather than record-by-record data). In two other cases, participants submitted datasets in code, which vastly increased the amount of time we had to spend in preparing datasets for publication. Because comprehensive data sharing is still relatively novel among this research community, the project leads felt the need to reduce barriers to participation (particularly time commitments). Thus, in order to motivate continued participation in this study, the editors attempted to make the project as undemanding as possible by taking on many of the data clean up and decoding burdens.

Coding systems still see widespread use in archaeology largely because they facilitate rapid data entry. As discussed, the zooarchaeological community has no standard set of codes, and each individually-coded dataset needs extensive

 doi:10.2218/ijdc.v9i1.301

documentation or decoding to be intelligible by the wider community. Because of the attention and expertise required to understand different coding systems, decoding datasets required the greatest amount of editorial effort. For example, one of the contributed datasets had over 125,000 specimens, but because it was decoded prior to submission, additional edits and preparation by Open Context's editors only required about four hours. In comparison, another dataset of only 15,000 specimens entirely in code took over 30 hours to translate. In other words, a coded dataset one tenth the size of a decoded dataset required ten times as much effort to prepare for use.

Over the course of this study, Open Context's editors devoted over 130 person-hours to these editorial steps. Domain knowledge on the part of the editor proved to be invaluable, allowing editors to distinguish trivial typographic errors from more serious errors or inconsistencies. Domain expertise was especially required for decoding datasets in preparation for shared analysis and publication. Decoding often revealed gaps in codebooks and other documentation, and their resolution required back-and-forth communication between data editors and data contributors. If the data contributors simply archived their datasets and documentation without editorial review, such documentation gaps would have likely gone unnoticed and unresolved.

The time and effort required in decoding data, together with risks of gaps in coding documentation, have important data management policy implications. Data management plans should explicitly address the issue of dataset coding because of costs and data quality concerns that will be faced by future users. Although decoding requires a great deal of effort, a decoded dataset never has to be decoded again. Thus, over the long term, decoding early in the lifecycle of data dissemination and archiving clearly saves a great deal of time and effort. One can apply similar logic to cleaning (editing) data.

Data Annotation

Aligning to ontologies is extremely useful for data integration because it disambiguates meaning and draws links between like terms that may have been recorded slightly differently. Though the application of community controlled vocabularies and ontologies for this project was straightforward, we encountered a few complications. For example, some classifications important to zooarchaeology lacked representation in the EOL or UBERON vocabularies. In the case of UBERON, the ontology needed a simple expansion. Open Context's editors requested new identifiers for skeletal elements that were missing from the UBERON ontology. The use of UBERON for annotating zooarchaeology datasets, thus, improved the coverage of UBERON to include elements that occur only in certain taxa.[8]

In other cases, an existing ontology may have related concepts, but those concepts may map poorly to a specific domain need. For example, zooarchaeologists frequently classify certain bone specimens as "sheep/goat" because of the difficulty in visually differentiating sheep from goat in the morphology of many bone elements. The project could have related classifications of "sheep/goat" to the EOL identifier for the "Caprinae"[9], a taxonomic subfamily grouping that includes sheep and goats. However, Caprinae also includes many taxa that zooarchaeologists would regard as highly improbable. Thus, we requested a new EOL identifier for the concept "sheep/goat".[10]

8 For example, the project requested the creation of the "fused tarsals 1 and 2" which occurs in equids. This term is now available in UBERON: http://purl.obolibrary.org/obo/UBERON_0013649

9 EOL identifier for "Caprinae": http://eol.org/pages/2851411

10 EOL identifier for "sheep/goat": http://eol.org/pages/32609438

We were able to collaborate easily with the managers of the EOL and UBERON vocabularies to extend the vocabularies as needed. Curators of ontologies and controlled vocabularies must be responsive to community needs and have processes to add new concepts as required. Without this flexibility to meet the needs of a particular domain, the vocabulary would be of minimal interdisciplinary applicability for data integration and linked data applications. In addition, the application of controlled vocabularies and ontologies relates to the decoding issue discussed above. In some cases, terminologies used in a specific dataset contained ambiguities. Though a sheep and a donkey both have a bone that may be identified as a "metacarpal", uses of the term "metacarpal" have ambiguities that may complicate future data reuse. In sheep, a "metacarpal" is more precisely defined with the UBERON concept of "fused metacarpal bones 3 and 4"[11], while a donkey "metacarpal" is (usually) more precisely the UBERON "metacarpal bone of digit 3"[12]. The more precise UBERON concepts better capture homology, developmental biology and evolutionary history of these bone elements. The editorial process of annotating a dataset with controlled vocabularies helps further resolve such ambiguities and document data in ways that can facilitate reuse.

Data Interpretation and Reuse

Most scientific data sharing and archiving efforts have a goal of opening new research opportunities. The effort and expense involved in the data publication processes of review, editing, documentation and annotation need to pay dividends in terms of compelling research outcomes. The participants in this project had confidence in using the edited and ontology-annotated data for many types of comparative analysis, particularly those forms of analysis less sensitive to sampling biases. However, certain forms of comparative analysis proved more challenging. Researchers needed more information about factors that may bias sampling. For example, some datasets in this study contained a large number of molluscs. Researchers needed to know if the absence of molluscs meant that the ancient inhabitants did not exploit marine resources, or that molluscs were simply not recorded in some databases. Understanding such "missing data" is critical for many forms of reuse and these types of sampling biases need documentation in the project metadata. Studies of data reuse in other domains note similar documentation needs (Faniel et al., 2012; Van House, 2002; Wallis et al., 2007).

Tooth data recorded by project participants proved very difficult to integrate and compare across contributing projects. Though all participants used the system for recording tooth eruption and wear developed by Payne (1973), the manner in which they recorded observations varied greatly. For example, one analyst noted the tooth number in the column heading ("Molar 1") and listed the tooth wear stage in the cell below. Another analyst noted the tooth number in a "Tooth Number" field and the wear stage in a "Wear Stage" field, while others recorded all tooth data in a "Comments" field. Though all used Payne's system, incompatibilities in organizing the tooth data, especially widespread reliance on free-text comments fields, made integration via an ontology too cumbersome to undertake in the context of this project.

These examples demonstrate how integrated analysis helps highlight areas where the discipline needs better data modeling practices. Traditionally, researchers present papers summarizing bone identifications from archaeological sites. Researchers shared few specifics about data management and modeling techniques. As we begin to look "under the hood" at datasets, the analytic importance of such techniques becomes more

11 UBERON identifier: http://purl.obolibrary.org/obo/UBERON_0013587
12 UBERON identifier: http://purl.obolibrary.org/obo/UBERON_0003647

 doi:10.2218/ijdc.v9i1.301

apparent. Recording practices with better analytic potential need to be adopted. Thus, the positive impacts of data dissemination not only apply to the data themselves, but also to better practices in data modeling and organization.

Discussion

Addressing these data reuse challenges requires improvements in practice at every stage in the management of data, from creation through dissemination and archiving. Thus, while data preservation alone motivates better data management, emphasizing *data reuse* as a professional goal will go even further to improve data management practices.

Improving Data Creation Practices

The data prepared for collaborative analysis in this study first passed through a process of editorial review, revision and annotation. As described above, this process involved significant effort and required domain knowledge of dedicated "data editors". Certain specific improvements in data management practices at the time of data creation can reduce downstream costs for both data editors and consumers.

- **Data Validation and Decoding:** Errors in coded data are difficult to notice, and coding documentation often does not exactly match coded data. Data in coded form (even when documented) greatly multiplied the effort required for reuse.
- **Better Data Modeling:** Poor data modeling can impede later data reuse, yet many researchers lack formal training in data management. Adequate modeling of complex phenomena, such as tooth eruption and wear, as simple, flat tabular data (such as a spreadsheet) is challenging and can impede data reuse.

In order to improve the prospects of data integration and reuse, the research community needs to adopt better data management techniques at the outset, at the stage of data creation (Faniel et al., 2013). Until researchers feel more rewards for reuse of their data, they will likely not invest more effort in improving data creation practices. Data management policy makers can improve this situation by recognizing the need for training and financing, and by rewarding new professional roles, particularly researchers that combine domain knowledge, data and software expertise. The new Institute for Data Science at UC Berkeley[13] represents an attempt to reorganize academic roles to better sustain data intensive research.

"Publishing" and "Pushing" Quality Research Data

As discussed above, datasets in small science fields like zooarchaeology do not magically come together to reveal new insights. The application and extension of controlled vocabularies and ontologies is necessary for their integration, and this requires effort and expert knowledge. Many datasets will require significant effort and domain expertise to be ready for reuse. The concept of "data sharing as publishing" helps to encapsulate and communicate the investment and skills needed for sharing reusable data. A publishing metaphor can help put that effort into a context that is

13 Berkeley Institute for Data Science at UC Berkeley: http://vcresearch.berkeley.edu/datascience/overview-berkeley-institute-for-data-science

recognizable by the research community (i.e. data publishing implies efforts and outcomes similar to conventional publishing). Offering a more formalized approach to data sharing can also help promote professional recognition (a key need noted by Harley, 2013), which would motivate better data creation practices at the outset. Ideally, "data sharing as publishing" can help create the reward structures that make data reuse less costly and more scientifically rewarding (Kansa & Kansa, 2013). This can help better situate data sharing within the Academy's conventions and traditions (see also Costello, 2009), perhaps complementing rewards through impact advantages sometimes observed on articles associated with open data (Piowar & Vision, 2013).

The "publishing" model and data quality

The process of data publishing presented here involves editorial review and revision processes that result in datasets of a higher quality than "raw" datasets. This raises an interesting question: what constitutes data quality? Data quality may largely depend upon the use put to data. For instance, all of the data published in Open Context are in some sense "reused". The datasets are taken from contributed databases indexed and displayed in Open Context's online interface. Problems in data and metadata quality can break certain functionality on the Open Context website. Open Context's editors, as well as its data contributors, have a motivation to promote data quality in order to build and maintain their professional reputations.

In this study, the quality of datasets and associated documentation relate to how well researchers could analyze and compare across datasets with confidence. In an important sense, this exercise in data integration and reuse provided a rigorous form of peer review. For example, questions around possible missing information (in the case of molluscs) presented a key stumbling block in comparative analyses. Many of the zooarchaeologists participating in this study sought additional information on sampling procedures and other factors that may bias representation of different species in these datasets. Did the absence of evidence imply evidence of absence? Reviewing the challenges researchers faced in reusing data for integrated analysis can inform our notions of data quality and lead to better documentation standards and editorial practices.

The "push" model of public version control

Some problems in data recording and documentation only became evident when researchers actually tried to reuse and analyze each others' datasets. The value of feedback from reuse helps to highlight some of the limitations of the metaphor of "data sharing as publishing". Problems in a dataset may go undetected, even after cycles of editorial review and revision. To some scientific communication reform advocates, "publishing" carries connotations of finality that impede collaboration. Datasets need not, and often should not, be fixed as static products. In some respects, datasets are like software source code, where they are usually expressed as structured text (like source code) and usually require a computer to use. Similarly, researchers may wish to revise datasets, adding records or new annotations to ontologies and controlled vocabularies, by "forking" them as developers "fork" (diverge) source code.

Thus, reform advocates have used the phrase "Don't publish, push!"[14] to capture the need to encourage greater dynamism and collaboration in scientific communication. As is the case with source code (where new versions are "pushed" to the community), version control systems can improve the management, professionalism and

14 Attributed to Jason Priem in a tweet by Carl Boettiger on 25th April 2013: https://twitter.com/cboettig/status/327534823830863872

 doi:10.2218/ijdc.v9i1.301

documentation associated with ongoing and collaborative revision of datasets (see also Winn, 2013). To help meet these needs, Open Context uses GitHub for dataset version control, including datasets involved in this project.[15] For example, the Çatalhöyük Zooarchaeology dataset involved contributions from over 34 zooarchaeologists. GitHub tracked several revisions of data and metadata to this large and complex dataset made in response to feedback about chronology and attribution metadata gathered from some researchers that did not directly participate in our EOL-funded study.

Mixing the models

In our view, elements of both the "publishing" and "push" models play a role in promoting better data management and reuse. The metaphor of "data publishing" encapsulates the effort and expertise involved in effective data sharing. At the same time, in order to not simply replicate the limitations of conventional publishing in data sharing, we should communicate the desirability of "pushing" revisions to published data. In the case presented here, data first passed through an internal process of editorial review and revision prior to public release into public version control systems (GitHub).

The relative emphasis on private editorial feedback and public version control will likely vary according to specific circumstances. In the context of this study, we wanted to offer an opportunity for researchers to gain private feedback on data before public disclosure. As the research community becomes more comfortable with data dissemination, editorial changes and revisions can be conducted more often in the context of public version control. This may offer more documentation and apparent "transparency" about a dataset. However, we should caution that "transparency" is a heavily loaded term that should be used with care. There are many factors that can shape data and its interpretive potential from the very beginning, even a dataset publicly available on GitHub. Version control systems like GitHub may lead to greater accountability and better documentation, and provide more opportunity for feedback and collaboration. However, version control systems will not automatically give "transparency" to research. Tacit biases and implicit sociological and cultural factors also shape data collection and research practices in ways that are not clearly evident in a context like GitHub. The impact of version control on data collection, documentation and reuse is an area that deserves further attention.

Regardless of the dissemination model (publishing, pushing, or simply uploading a minimally documented spreadsheet into a digital archive), any communication of data is goal-directed. Researchers will use these various dissemination models to demonstrate and publicize their accomplishments, collaborate with colleagues, comply with rules imposed by funding agencies, or meet other goals. In every case, they will select exactly which data to share, when to share it, and to what level of detail and quality. Thus, though data dissemination can indeed open new research opportunities, particularly in areas involving data integration, scientific objectivity will not automatically emerge from these practices.

Conclusions

This study highlights the importance of regarding data as more than a "residue" of research that should be archived. Our case study shows that data sharing and reuse, even among a group of specialists in a small sub-discipline, can be an extremely complex

15 See example: https://github.com/ekansa/opencontext-eol-zooarch

process, often involving a great investment in adequately preparing and documenting data. Value-added models of data dissemination that involve much more careful scrutiny in the form of editorial review and analytic reuse by outside researchers, can improve the reusability of data. In addition, this study highlights the importance of studying the challenges researchers face when attempting to reuse each other's data. It is very difficult to anticipate all the data documentation that may be required to inform future reuse without more experience of data reuse in practice.

In our case study, we first followed a "publishing" model where dedicated editors assisted in data clean up and documentation. After this initial stage, additional refinements on the datasets took place in using a "push" model of public version control (GitHub), in a manner similar to software debugging and issue tracking. The experiences documented in this case study help illustrate how "publication" and "push" models of data dissemination need not be mutually exclusive; on the contrary, they can play complementary roles in sharing high quality data in support of research.

More experience with data reuse will inform ways to better shape professional practices around data dissemination. Data often need to go through cycles of vetting, review, revision, annotation and documentation by collaborating editors and/or peers. Without processes of collaborative coproduction (including "publishing" and "push" models of dissemination), many datasets will not be useful in the future. Models that mix the formalism of publication (in terms of dedicated expertise and professionalism) with the continual feedback and revision cycles of public version control can improve data dissemination practices.

Acknowledgements

The authors would like to thank the Encyclopedia of Life for enabling this research through a Computable Data Challenge award to the Alexandria Archive Institute for the *Biogeography of Animal Domestication Using EOL* project. This research was also supported in part by a Digital Humanities Implementation grant from the National Endowment for the Humanities (award HK-50037-12). Any views, findings, conclusions, or recommendations expressed in this work do not necessarily reflect those of the Encyclopedia of Life or the National Endowment for the Humanities. The authors would also like to thank Cheryl Makarewicz (Institute of Pre- and Protohistoric Archaeology, CAU Kiel) for facilitating the conference session that brought together the project participants at the International Open Workshop, April 16-19, 2013, Kiel University. Finally, the authors would like to recognize this project's many participants, whose patience, enthusiasm, and scholarly contributions are much appreciated.

References

Arbuckle, B.S., Kansa, S., Kansa, E., Orton, D., Çakırlar, C., Gourichon, L., Atici, L., … , Würtenberger, D. (2014). Data sharing reveals complexity in the westward spread of domestic animals across neolithic Turkey. *PLoS ONE, 9*(6), e99845. doi:10.1371/journal.pone.0099845

Costello, M.J. (2009). Motivating online publication of data. *BioScience, 59,* 418–27. doi:10.1525/bio.2009.59.5.9

Dibble, H.L. & McPherron, P. (1988). On the computerization of archaeological projects. *Journal of Field Archaeology, 15*(4), 431–440. doi:10.1179/jfa.1988.15.4.431

Eiteljorg II, H. (1998). Archiving archeological data in the next millennium. *Cultural Resource Management, 21*(6), 21–23. Retrieved from http://crm.cr.nps.gov/archive/21-6/21-6-6.pdf

Faniel, I.M., Kriesberg, A., & Yakel, E. (2012). Data reuse and sensemaking among novice social scientists. In *Proceedings of the American Society for Information Science and Technology, 49*(1). doi:10.1002/meet.14504901068

Faniel, I.M., Kansa, E.C., Kansa, S.W., Barrera-Gomez, J. & Yakel, E. (2013). The challenges of digging data: A study of context in archaeological data reuse. *JCDL 2013 Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries.* New York, NY: ACM.

Griffiths, A. (2009). The publication of research data: Researcher attitudes and behaviour. *International Journal of Digital Curation 4*(1), 46–56. doi:10.2218/ijdc.v4i1.77

Harley, D., Acord, S.K., Earl-Novell, S., Lawrence, S. & King, C.J. (2010). *Assessing the future landscape of scholarly communication: An exploration of faculty values and needs in seven disciplines*. Berkeley, CA: Center for Studies in Higher Education.

Harley, D. (2013). Scholarly communication: Cultural contexts, evolving models. *Science, 342,* 80–82. doi:10.1126/science.1243622

Kansa, E. (2012). Openness and archaeology's information ecosystem. *World Archaeology, 44*(4), 498–520. doi:10.1080/00438243.2012.737575

Kansa, E.C., & Kansa, S.W. (2013). We all know that a 14 is a sheep: Data publication and professionalism in archaeological communication. *Journal of Eastern Mediterranean Archaeology and Heritage Studies, 1*(1), 88–97. doi:10.5325/jeasmedarcherstu.1.1.0088

Onsrud, H.J. & Campbell, J. (2007). Big opportunities in access to "small science" data. *CODATA Data Science Journal*, *6,* OD58-OD66. doi:10.2481/dsj.6.OD58

Payne, S. (1973). Kill-off patterns in sheep and goats: The mandibles from Aşvan Kale. *Anatolian Studies, 23*, 281–303. doi:10.2307/3642547

Piowar, H. & Vision, T.J. (2013). Data reuse and the open data citation advantage. *PeerJ PrePrints*. doi:10.7717/peerj.175

Van House, N. (2002). Digital libraries and practices of trust: Networked biodiversity information. *Social Epistemology: A Journal of Knowledge, Culture and Policy, 16*(1), 99. doi:10.1080/02691720210132833

Wallis, J.C., Borgman, C.L., Mayernik, M.S., Pepe, A., Ramanathan, N., Hansen, M. (2007). Know thy sensor: Trust, data quality, and data integrity in scientific digital libraries. In *European Conference on Research and Advanced Technology for Digital Libraries, 4675,* 380–391. Budapest, Hungary. doi:10.1007/978-3-540-74851-9_32

Wallis J.C., Rolando, E. & Borgman, C.L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS ONE, 8*(7), e67332. doi:10.1371/journal.pone.0067332

Winn, J. (2013). Open data and the academy: An evaluation of CKAN for research data management. Paper presented at the IASSIST 2013, Cologne. Retrieved from http://eprints.lincoln.ac.uk/9778/