SPECIAL FEATURE: PERSPECTIVE

Promoting data quality and reuse in archaeology through collaborative identifier practices

Eric C. Kansa[a,1] and Sarah Whitcher Kansa[a]

Edited by Suzanne Pilaar Birch, Anthropology and Geography, The University of Georgia, Athens, GA; received August 2, 2021; accepted October 29, 2021 by Editorial Board Member Dolores R. Piperno

Investments in data management infrastructure often seek to catalyze new research outcomes based on the reuse of research data. To achieve the goals of these investments, we need to better understand how data creation and data quality concerns shape the potential reuse of data. The primary audience for this paper centers on scientific domain specialists that create and (re)use datasets documenting archaeological materials. This paper discusses practices that promote data quality in support of more open-ended reuse of data beyond the immediate needs of the creators. We argue that identifier practices play a key, but poorly recognized, role in promoting data quality and reusability. We use specific archaeological examples to demonstrate how the use of globally unique and persistent identifiers can communicate aspects of context, avoid errors and misinterpretations, and facilitate integration and reuse. We then discuss the responsibility of data creators and data reusers to employ identifiers to better maintain the contextual integrity of data, including professional, social, and ethical dimensions.

identifiers | data curation | collaboration | archaeological science | interoperability

Archaeological domain specialists, such as zooarchaeologists, paleoethnobotanists, archaeometallurgists, and certain ceramic specialists, often have research interests and recording systems that borrow from or overlap with disciplinary communities outside of archaeology. In contrast with excavation or survey directors, who may invest years leading a single project, many domain specialists are "itinerant," participating in multiple archaeological projects and studying some portion of the collected materials on a short-term basis. They often have research needs that require creation and use of datasets from multiple archaeological excavation and survey projects. This paper aims to promote feasible good practices among domain specialists that study and document archaeological materials to make data better service integration and aggregation required for these research needs. The need for such guidance is clear. Archaeology often employs destructive methods; therefore, it is imperative for archaeologists to create, manage, disseminate, and preserve detailed and high-quality records of their observations. Increasingly, observational records are in the form of various types of digital data and media. While digital recording (often alongside paper recording) is ubiquitous in the discipline, current normative practices do not necessarily promote data preservation, reuse, or understanding (1). Poor data management practices (how data are created and curated) may lead to information loss, and undetected errors in data can cause confusion and lead to misinterpretations.

Although archaeologists may recognize that more thought and effort need to be put into good data curation practices that support data preservation and access (2), they have few financial resources and limited access to technical expertise and information services to support their specific needs. Poor data management practices may further exacerbate these problems. Problems associated with poorly conceived or executed data creation can compound and become time consuming and costly to resolve later, if at all. Therefore, investments in data preservation and archiving need to be accompanied by investments that promote greater professional recognition and impact for creating and sharing high-quality data. Practices that improve quality at the point of data creation would facilitate data reuse

[a]Alexandria Archive Institute, Open Context, San Francisco, CA 94127

Author contributions: E.C.K. and S.W.K. designed research; E.C.K. and S.W.K. performed research; and E.C.K. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. S.P.B. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: eric@opencontext.org.

Published October 17, 2022.

and promise to multiply the research and instructional impacts of investments in data archiving (3, 4).

This paper argues that identifier management practices play a key role in the choreography of data created and used by multiple specialists with different areas of disciplinary expertise. Data need to flow efficiently and accurately among these specialists. Without good identifier management, specialist datasets become "siloed," thereby limiting opportunities for synthetic understandings based on the inputs of multiple specialist studies. This paper describes ways domain specialists can move from siloed (isolated) data management practices toward more collaborative practices that create data suitable for reuse in support of a broader range of research questions and applications. Better identifier practices can promote broader data reuse and interoperability both within archaeology and between archaeology and adjacent scientific disciplines.

Background

Data practices are receiving increasing attention in the context of data management plans (DMPs) now required by many granting bodies. However, DMPs by themselves do little to encourage creation of data that see wider reuse by the research community. In a study of 834 DMPs from an Australian university, Smale et al. (5) found that the types of project-wide DMPs required by many funders have little impact on data reuse, highlighting the need to change perceptions about the purpose of DMPs. Rather than focusing on compliance aspects, proponents of good research data practices should highlight how the creation of quality, reusable data can play a positive role in addressing an investigator's central research goals.

The research goals of specialists often require creation and reuse of datasets documenting materials from multiple excavation and survey projects. The Secret Life of Data (SLO-data) project (3, 6) explored how these goals motivate data management practices orthogonal to the goals of excavation directors who invest far more time and focus on a single project. Drawing on observations and interviews of researchers on field projects, the project identified several common areas of friction in data management across the participating excavations (3, 6) that centered on communication and coordination among team members, and especially domain specialists.

Preparing data for publication and integration requires significant clean up and translation work (7–9). Adapting complex data "late in the game" can be costly, and some of those costs could be avoided if teams aligned data creation practices for the needs of wider community reuse from the start. However, many problems in data modeling (organization and layout) at the start of a project impact data quality and only become apparent in later data reuse. Gaining experience in reusing data created by others can build expectations about what constitutes "good data." Data reusers can apply their experiences to become better data creators and help jumpstart virtuous cycles (4) of better data creation that encourage more and higher-impact reuses of data.

Identifiers play a fundamental role in shaping data quality and reusability. Good data management practices also involve data modeling, validation practices to promote consistency and reduce errors, specifying data type (enforcing integer, decimal, or Boolean values, where needed), use of open and nonproprietary file formats, and much more (10, 11). Geospatial data (12, 13), archaeological survey data (14), zooarchaeological data (15), and archaeobotany (16) all have their own more specialized data quality concerns. Much of this prior work focuses on aspects of quality for datasets viewed as relatively discrete and isolated works. In contrast, this paper's focus on identifiers emphasizes a more relational and contextual view of data quality.

Context plays a central role in archaeological method and theory. Archaeologists use documentation of contextual relationships to interpret observations of features, artifacts, and ecofacts, individually and in aggregate. The centrality of context in archaeological practice makes archaeological data highly "relational" in that each individual dataset often references entities described elsewhere in other datasets. We consider the capacity of datasets to support analytically useful connections to other related information that may be key to interpretation and analysis as "contextual integrity." For this reason, a major factor shaping the quality of archaeological data centers on the ability to efficiently and reliably look up and access data documenting related entities that can inform archaeological understanding of context.

Data Quality: General Aspects. "Data quality" is surprisingly difficult to precisely define or specify. Most data management experts define "high-quality" data as data that satisfies needs for intended uses. For example, Wang et al. (ref. 17, p. 625) describe data quality in terms of utility. Wang and Strong (18) emphasize the value of an empirical approach to assessing data quality because it captures the voices of data consumers in judging the fitness of data for reuse, rather than theoretical or intuitive judgements. This shifts the focus away from data development to its use and helps distill which aspects of data quality are most important to consumers. Cai and Zhu (19) elaborated on earlier "fitness for use" conceptualizations of data quality to consider more contextual dimensions, especially metadata, documentation, and credibility concerns.

Because quality of data relates to its intended use, data quality concerns become more complex, fluid, and difficult to specify if intended uses change or are unknown. A dataset adequate to meet the needs of an individual specialist working in isolation may be problematic for use by future researchers and future research programs. However, one of the major goals of research data management is to encourage new discoveries through the reuse of research data shared among members of wider communities. As more public and open datasets become available, researchers will have more reason to attempt to understand and assess data they did not create themselves. To realize the potential of open data, Sadiq and Indulska (20) stress the need for better means of assessing data quality more generically (that is, without intended use predetermined). Data users "need to be empowered with exploratory capabilities that will allow them to investigate the quality of the data sets and, subsequently, the implications of their use" (ref. 20, p. 152).

Data Quality in Siloed vs. Collaborative Contexts. Increasing expectations for data sharing and archiving require that our criteria for understanding data quality need to evolve. If part of the point of preserving data is to encourage reuse for new applications, then expectations about data quality need to better align to how well data can service multiple and even open-ended new uses. In archaeology, research necessarily draws on datasets created by many different people—from context information documented during excavations to specialist analyses on macroscopic remains to chemical analyses conducted in

laboratories often years after the excavation occurred. It is critical that these various data types are linked and intelligible by others. Identifiers enable this linking and promote data quality that supports more open-ended reuse of data beyond the immediate needs of the creators.

A key aim of this paper is to identify practices around identifier management to help individual researchers and teams prepare data for future reuse. Good identifier management helps address the common data challenges of completeness, aggregation, and specificity—attributes that help shape opportunities for future data reuse. Data reuse can take many forms. In some cases, "reproducible research" goals of reuse would emphasize attempts to use data to replicate analytic claims. Other uses of research data include instruction and professional development, where a dataset germane to a research topic or area of disciplinary interest would be used for training and teaching purposes. In other cases, a dataset would be "sampled" or aggregated in some fashion to be combined with data from other sources. This can be done on a record-by-record level (citing specific comparanda) or by aggregating data to some level of granularity convenient for analytic purposes. For example, a researcher may want to compare summary statistics of the relative frequency of different biological taxa in zooarchaeological assemblages from different sites, and they may not need to reference individual bone specimens. But other studies that may investigate patterns in the butchery of different body parts may require access to records documenting specific cut marks on individual bone specimens.

Thus, in considering uses of data beyond reproducibility studies, questions of aggregation and completeness become more central. A dataset need be only as specific or complete as required to support the analytic claims in a publication for it to be suitable for a replication study (see discussion of reproducibility in archaeology in refs. 21, 22). However, aggregated data reported as such cannot be disaggregated. Data reported at higher levels of granularity offer opportunities for different kinds of aggregate analysis. If data quality is a reflection of the suitability of a dataset for certain applications, then more granular and specific data are of higher quality because these data offer a greater range of options for future analyses. Therefore, good data practices must consider needs beyond what may be required to support replication of a single isolated study.

Quality and Completeness Issues in Siloed Contexts. Research that operates in isolation to address discrete questions can lead to overly siloed data creation practices. Siloed data practices can lead to collecting and reporting data at a level of aggregation that does not offer enough granularity and specificity to support a wide range of analytic options. Similarly, siloed data practices may create data in insolation, with few if any connections or potential connections to any other dataset. Such isolation can undermine the "completeness" of a dataset and negatively impact future analytic opportunities.

Data can be considered more complete if it is richly described and linked to contextualizing datasets. Unfortunately, many common workflows involving domain specialists make data reporting less than complete. Commonplace siloed practices can isolate domain specialist data creation with little broader coordination. Even within a given project, individual specialists may create their own datasets independently without access to data curated by other project team members. Many projects lack clear expectations or processes to bridge across the siloed datasets created by team members, making data integration even within a given project a challenge.

Many of the datasets published by Open Context reflect these practices. Open Context may publish a zooarchaeological dataset from a project, but this dataset would lack detailed context records or other records documenting material culture and other observations. This pattern reflects how different areas of professional expertise have their own publishing cadences and research interests. These siloed practices typically lead to incomplete and piecemeal data dissemination. Potentially useful related data (that would augment completeness and, by inference, quality) may be absent or may be hard to integrate even if they eventually become available. As discussed below, bridging and improving upon these siloed practices requires greater attention to identifier management.

Identifier Practices and Data Quality

Archaeologists use identifiers to name and reference individual archaeological sites, stratigraphic units, features, objects, ecofacts, samples, permits, and even people, reports, and publications. As such, identifiers play a key role in all of the relational aspects of data (23, 24). This is true even if data are not stored in a relational database! In common practice, a given excavation project may have multiple databases managed by different domain specialists. However, identifiers to archaeological contexts can often be inconsistent across these datasets (Fig. 1*A*). Domain specialists, working independently and often transcribing shorthand from paper specimen tags frequently create data that have such identifier mismatches (10).

Because identifiers are key to linking and cross-referencing related data, they play a central role in expressing and communicating many aspects of context. Breakdowns in the management and use of identifiers undermine the contextual integrity of data. In the example above, misalignment of context identifiers in the zooarchaeological dataset and the context database break contextual associations of the zooarchaeological observations. Unrecognized errors in contextual associations may lead to mistaken inferences, as records from different datasets (for instance, a zooarchaeological and a pottery dataset) may get associated in error. Similarly, unnoticed, repeated use of the same identifiers may lead to miscounting of individual specimens.

Toward Collaborative Identifier Practices. Many archaeological projects face difficulties in reconciling identifiers across datasets created by different collaborators (3, 24). Rather than expecting deeply entrenched practices to change overnight, archaeologists need more feasible pathways toward incrementally adopting better practices. Fig. 1 *A–C* outline readily achievable improvements in practice that can be adapted in the context of individual research, in the context of a small team, and in more global contexts with publicly curated data.

A key aspect of improving identifier practices requires that data creators consider what parts of their data may relate to entities described in other datasets. In database jargon, this involves making distinctions between "literals" and "identifiers." Not everything in a dataset is an identifier. A zooarchaeological data table may include a column "GL" (meaning "greatest length") with length measurements in millimeters. The numbers in the GL column are literals, meaning they are simply values; they are not tokens for anything else. The same data table can have a numeric "locus" column. But in the case of the locus column, the values

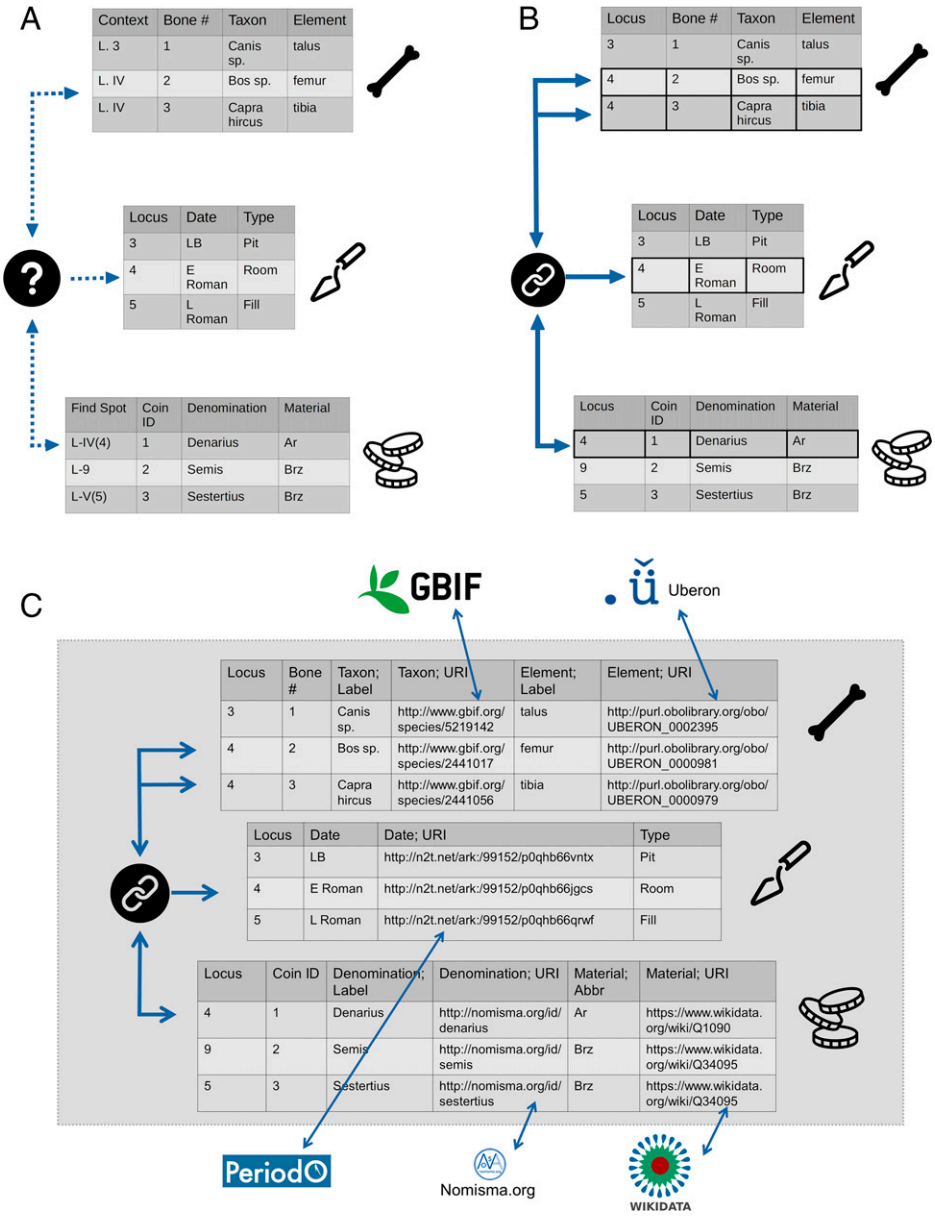

Fig. 1. (*A*) Zooarchaeological and numismatic examples of siloed and isolated identifier practices (poor context linking). (*B*) Zooarchaeological and numismatic examples of team collaborative identifier practices (linking context IDs). (*C*) Zooarchaeological and numismatic examples of global collaborative identifier practices (linking contexts and publicly curated resources).

may be identifiers if they name entities described elsewhere. In siloed data practices, most data elements act as literals that cannot reliably reference outside information. To improve practice, teams of researchers should coordinate and agree upon what data elements need to be treated as identifiers (and expressed according to common rules and conventions) and what data elements can be treated as simply descriptive values (literals).

Projects that involve multiple specialists should routinely attempt to integrate specialist data so that problems can be diagnosed and resolved early. This requires project members to set common expectations and coordinate more frequently and deeply, as recommended by Faniel et al. (3). Clear documentation conventions and validation practices to check identifier associations across multiple project datasets will promote greater contextual integrity.

Good Identifiers for Global Collaboration

Identifiers can play key roles in shaping data quality and reuse more broadly, outside the scope of individual research projects. Good identifier practice must satisfy the following criteria:

- Actionable: Ideally, one must be able to look up and access ("resolve") at least some information about the entity named by an identifier. In Fig. 1*B*, the locus 4 identifier is actionable, because it is consistently expressed, and both the zooarchaeology and the coins tables directly associate with locus 4 in the context description table. A look-up does not require full open access to information about a named entity. In some scenarios, an identifier may point to sensitive data with access restrictions, but learning that such sensitivities exist also provides important contextual insights even if one does not have permission to access the details of a record.

- Unambiguous: A good identifier must uniquely name a specific item. Identifiers that are repeatedly used for multiple items can cause confusion.
- Shared and reused: Using and reusing shared identifiers is a key aspect of linking and cross-referencing data in scenarios of collaborative data creation and curation. This collaboration can take place both within a given research project, or asynchronously across multiple projects where researchers cross-reference data they create with identifiers in datasets created earlier by other teams.

In practice, a small team of researchers working within a specific project can agree upon conventions and rules to make identifiers that are good (according to the criteria defined above) within the local context of their specific project. Good local identifier practices are typically a necessary first step toward good global identifier practices and help smooth the way for publicly releasing data with greater quality and research value.

Researchers do not need specialized infrastructure or software to do a good job managing identifiers within their own projects. Data creation workflows that include coordination, checks, and enforcement of locally defined conventions for project participants are the main requirements. However, in wider contexts, good identifier management needs the support of dedicated infrastructure (25). Publishers of "linked open data" try to guarantee the persistence of URIs (Web addresses intended to also serve as identifiers) they assign to the various ontologies and data records they publish. Infrastructure services commit to issuing identifiers such as archival resource keys (ARKs) and document object identifiers (DOIs) that are long lasting (persistent) and uniquely resolve to specific resources curated by a given collection (26). ARKs and DOIs are both actionable in that they are typically supported by "resolution services" to look up information about an identifier on the World Wide Web. Globally unique identifiers are better suited for data aggregation and integration because, by definition, any given globally unique identifier will not be (mistakenly) repeated in different data sources.

Many projects define their own local identifiers for different entities like "feature 10," "locus 101," and "bone 124." These local identifiers should be unique within the context of a given project. While such practices are perfectly appropriate within individual projects, researchers must be careful in managing local identifiers when integrating data from multiple sources. Different projects and data sources will likely have used the same name, bone 124, for different specimens. Thus, in data aggregation and integration scenarios, globally unique identifiers become necessary.

Persistence, as well as uniqueness, is also an important concern for identifiers when collaborating or integrating data more broadly. Data creation often takes place asynchronously, sometimes over many years. For example, in 2006, Open Context (27) published a dataset of mammals and birds from Epipaleolithic contexts at Pınarbaşı, Turkey. Fourteen years later, a second dataset documented bird bones from the same site (28). Some records in the second dataset documented specimens already published in the 2006 dataset. Fortunately, the author was able to cross-reference records with the identifiers in the original dataset. If she had not been able to do this, many of the bird specimens would have been published twice, which would falsely inflate the representation of birds in the dataset, leading to potential misinterpretations about ancient environments, diet, and economy. Thus, identifier management is a concern that goes beyond the scope of an individual project. New studies may create additional data that relate to content already documented and published in prior investigations. Thus, identifiers need to be both globally unique and persistent.

To be effective, globally unique and persistent identifiers need to have both an analytically meaningful level of granularity, as well as supporting services and infrastructure. Globally unique and persistent identifiers can be minted at a very granular level, for each specific "entity" (bone, sherd, context, site, classification concept, etc.) described in archaeological datasets. Open Context has published and minted (through the EZID service) persistent, globally unique identifiers at a high level of granularity (for individual site, context, ecofact, artifact, and other records). ARK identifiers comprise the majority of the persistent identifiers minted by Open Context, and any other system can also mint ARK identifiers free of charge (for example, the Smithsonian, the Louvre, Institut de l'Information Scientifique et Technique, and many other institutions also maintain ARK identifiers for their archaeologically relevant collections). This high level of granularity in the application of persistent globally unique identifiers enables researchers, over the long term, to precisely and unambiguously cite, cross-reference, and maintain contextual associations of specific records across multiple datasets. Examples of cross-referencing include tDAR and Open Context (that link together archaeological site identifiers), VertNet and Open Context (for zooarchaeological specimens records, see ref. 29), and Pelagios (that links several archaeological and museum collection databases using shared identifiers to geographic places defined by public gazetteers).

However, most archaeological data repositories do not provide persistent identifiers at this level of granularity. DOIs, a type of persistent identifier most widely used by data repositories in archaeology, typically get assigned to entire datasets (see ref. 30), not to the individual entities documented by these datasets. A spreadsheet or relational database in an archaeological repository may describe thousands of individual archaeological contexts, artifacts, and ecofacts, but that entire dataset may have only one DOI.

Networking Identifiers and Archaeology's Grand Challenges

Good identifier practices encourage researchers to create data that link to other related data and can, in turn, be linked and cross-referenced with datasets created by others in the future. As such, good identifier practices are prerequisites for addressing "grand challenges" (2, 31) in archaeological research that require the integration and synthesis of large-scale datasets.

Data integration is often conceptualized as a process of networking information from heterogeneous sources (32). Many of the key supporting technologies behind data integration programs use "graph" data models to describe how instances of data network together with shared vocabularies and ontologies. Since "networking" is a widely used conceptualization for data integration, we should consider how to encourage the growth of scientifically valuable networks of research data. In the 1980s a telecommunications engineer named Bob Metcalfe (see ref. 33) observed how network connectivity grew with the square of the number of nodes on the network. Simply put, if you were the sole owner of a telephone, this network device would be useless since you would have nobody to call. However, as more

devices connected to the telephone network, each telephone gained more and more potential connections, presumably reflecting the growing value and utility of the entire network.

This observation can also apply to research data (see refs. 34 and 35 for the biological sciences). Schultes et al. (36) see parallels between the adoption of common Internet network protocols from the 1970s to the 1990s and current efforts to promote open research data practices, particularly the FAIR (findable, accessible, interoperable, reusable principles; see below). Persistent identifiers form the bedrock for linking elements of data into these integrative networks. However, persistent identifiers applied solely to an entire dataset (typical of normal repository practices, see above) leave only one point of potential connection. If a dataset contains many different persistent identifiers to concepts and other entities of interest (archaeological sites, coins, pottery types, sculpture, etc.), then that single dataset would vastly expand the scale and flexibility of potential networking. If "Metcalfe's Law" applies to research data, then wider and more granular use of persistent identifiers that network together data would multiply the impact and value of our investments in curating those data. The public reuse of persistent identifiers can fuel asynchronous and decentralized collaboration in building an integrative data network.

Feasible and Collaborative Practices for Catalyzing Network Effects. One of the challenges in catalyzing network effects is that most research data repositories are oriented around identification and discovery services of data files, not the individual entities described within those data files. While Open Context and a few other specialized data servers support persistent identification and discovery of records with more granularity, they may not be suitable for every archaeological dataset or research application. Ideally, an individual researcher should be able to mint persistent ARK identifiers and assign these to more granular entities. The Arketype.ch will soon enable any individual researcher to assign persistent identifiers at any desired level of specificity. In parallel, digital repositories should prioritize support for more granular referencing to specific records. Such specificity in identification can help make data integration more transparent and reproducible since the provenance of all data elements aggregated into a larger combined dataset can be precisely specified.

Using persistent identifiers does not require any specialized data management skills. Scientific domain specialists working with archaeological collections can and should create and reuse persistent identifiers using tools no more sophisticated than spreadsheets, relational databases, and Web browsers (see examples in ref. 10). Datasets that contain identifiers to ontology or controlled vocabulary concepts, to comparative specimens in museum collections, or to archaeological sites and time periods (especially with PeriodO), all can augment the intelligibility and interoperability of a dataset. This is illustrated in Fig. 1*C*, where a zooarchaeologist who adds columns in their spreadsheet with links to Global Biodiversity Information Facility (GBIF) (biological taxonomy), Uberon (anatomy), and PeriodO (chronology) can vastly enhance the clarity of their data. A dataset containing such identifiers can be more easily aggregated with other datasets containing these identifiers, thereby enabling integration of taxonomic, anatomical, and chronological observations (9). Perhaps more significantly, use of GBIF, Uberon, and PeriodO identifiers may also enable greater multidisciplinary reuse of data, since these infrastructures are in use by multiple information systems serving multiple disciplines.

While individual domain specialists in archaeology can already productively use persistent identifiers, better discovery services and other infrastructure would greatly augment the research value of shared identifiers (25). The newly launched "Internet of Samples" (iSamples) project aims to deliver such needed infrastructure to make the discovery and management of globally unique persistent identifiers for physical samples (including archaeological artifacts and ecofacts) useful for research (37). With better supporting infrastructure and the participation of major archaeological and natural science collections and information systems, the research value of such well-managed persistent identifiers will be easier to demonstrate and promote.

Discussion: Collaboration and Context

We have highlighted that high levels of granularity and good identifier practices can catalyze "network effects" to enhance the research impact and value of data sharing. Over the long term, these practices would make publicly archived and shared research data better support the kinds of large-scale, synthetic, grand challenge research agendas. However, reaching a critical mass of highly networked interoperable data can require long-term investments (38). Therefore, it is also important to define more incremental and feasible positive steps that researchers can take toward that vision. What more immediate and near-term goals can domain specialists achieve by moving from siloed toward more collaborative processes (Fig. 2)?

When archaeologists consider "context," they usually see stratigraphic deposits or survey locales as the main elements of

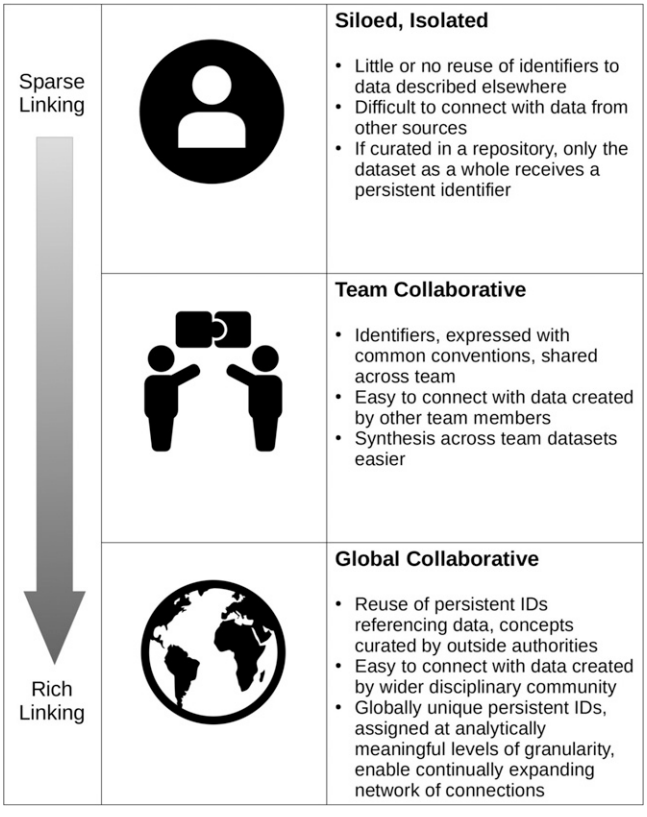

Fig. 2. Overview of siloed-to-global collaborative practices.

contextual information. However, context can include intellectual and conceptual dimensions beyond stratigraphic provenance. Examples include:

- identification of calibration targets and standards used in neutron activation analysis or X-ray fluorescence;
- identification of specific specimens used from a reference collection (and links to them, if available), such as for zooarchaeological or paleobotany documentation;
- identification of the specific instruments and or laboratories that generated results;
- identification of specific software application and library version information used to process or analyze results;
- identification of concepts (and links to them, if possible), including those defined in professionally curated controlled vocabularies or ontologies that can help align certain dimensions in a dataset with shared standards;
- links to comparanda, such as objects in museums or datasets created by other projects can help enrich contextual understanding; and
- links to ORCIDs for all individuals who participated on a project so data reusers can get a better sense of the research expertise on the team.

These benefits will not be realized if identifiers for people, concepts, calibration targets, comparanda, etc. cannot be used to look up and access the subject of those identifiers. Instead, persistent and globally unique identifiers are better suited to communicate these elements of analytic context.

Encouraging Changes from Siloed toward Collaborative Practices. Common use of shared identifiers promotes new capabilities for seeing patterns in data. As discussed, use of common identifiers for shared concepts in biological taxonomy and anatomical elements enables forms of data integration with demonstrated research impacts in zooarchaeology (9, 29, 39). Reference to outside standards does not limit innovation in recording systems; rather, identifiers that reference publicly curated standards can coexist with a researcher's own custom-defined data attributes. This helps to bridge specifically tailored aspects of a dataset with aspects shared by wider disciplinary communities. Motivations to make identified shared concepts include:

- Reduced data documentation burden. A dataset that uses persistent identifiers to reference concepts and entities already curated by external authorities is at least partially "self-describing," meaning that a researcher can spend less time and effort documenting every specific attribute of their own dataset since that documentation is already available from an external authority. That time can be better spent devoted to documenting the unique and custom attributes of their own data. Thus, the reuse of persistent identifiers can help reduce the labor and costs of metadata creation.
- FAIR data impacts and recognition. Use of common identifiers for shared concepts helps align datasets with the FAIR principles (40). Such identifiers can be widely discoverable, and they promote greater interoperability and reusability. Presumably, these characteristics should make the actual reuse of research data more likely. Capture of data citations and metrics can demonstrate research impact, leading to greater professional recognition and rewards for the creators of the original data.

Data integration through common reference to identifiers is not limited to concepts in ontologies or controlled vocabularies. Museum and other reference collections can also be used. For example, if certain specimen records in different zooarchaeological datasets all reference an identifier to a comparative specimen documenting a skeletal pathology, then researchers may see opportunities to use these data in follow-up paleopathology studies.

New research applications, reduced labor in data documentation, and potential professional rewards for creating higher-impact, more reusable data may all offer positive incentives for individual researchers to shift from siloed toward more collaborative data curation practices. This shift also needs to be supported by professional societies and granting bodies (especially reviewers of DMPs) and professional expectations for data sharing, use of data repositories, and publishing with data journal services (such as the *Journal of Open Archaeology Data*). Data sharing currently often takes place through the dissemination of "supplemental information" (SI) associated with peer-reviewed articles. However, in cases where data are shared as SI, there is little incentive to contextualize data using shared identifier systems or adopt other aspects of FAIR data principles (see ref. 41 for a review of compliance issues with journal data policies). Professional reward systems need to treat data as more than an adjunct or supplemental outcome of publishing articles (9). Peer review processes, both for publications and grant awards, need to do more than simply encourage data sharing; they need to encourage (and reward) data curation practices that promote broader contextualization and wider reuse of data.

Identifiers and Evolving Context. The potential associations one can make between datasets change over time, typically growing as network effects expand opportunities for connection. Nomisma.org (a database supporting domain specialist numismatic research) aggregates data records from dozens of (FAIR data practicing) online collections, including Open Context. Nomisma.org applied its own ontologies and controlled vocabularies to recontextualize and better harmonize these legacy datasets according to standards used by these domain specialists. Identifiers provide a means to more precisely and explicitly demonstrate provenance while defining these new contextual frameworks.

The example of Nomisma.org highlights one key advantage of using persistent identifiers to reference specific records of data, rather than more oblique citation of data described by published literature. Typically published literature discusses data in aggregate, so citation of a publication about a given dataset lacks granularity or specificity. In the case of Nomisma.org, the recontextualization of precisely identified records of data are also "machine actionable." Nomisma.org expresses its own (re)organization of data using open data formats and explicit ontologies, all easy to access and process with widely used open source software. Such computational recontextualization makes tracking of data provenance and reuse with new conceptual models (especially ontologies) both explicit and feasible at large scales.

Identifiers, Contextual Integrity, and Ethics. In promoting contextual integrity, identifier management is not merely a technical consideration, but it can play an important role in professional conduct and ethics. Data creators have a responsibility to document professional, social, and ethical aspects of context in their data. This may include:

- Identify roles, responsibilities, and recognition expectations for different people involved in the creation of a dataset. This

includes identification and recognition for data creators and different people involved in different steps of analyses. Common metadata for documenting roles as well as persistent and globally unique identifier services like ORCID.org can help communicate professional roles in data creation.

- Identification of social expectations extends beyond academic contexts. It also includes recognition for the roles and expectations of different stakeholder communities (especially indigenous or other descendent communities). Such recognition can be formally expressed in datasets using frameworks like Local-Contexts.org (37).

While data creators have responsibilities to communicate these professional and social aspects of context, data reusers also have responsibilities to respect and maintain the contextual integrity of reused data. Identifiers that help communicate social and professional context express key aspects of data provenance and some of the conditions and expectations of the parties involved in data creation. Therefore, data reusers should practice the following:

- Maintain identifiers of data they cite and sample. Those identifiers help describe social context of data (who created it, who may be stakeholders, specific expectations of stakeholders).
- Reference specific agreements, expectations, acknowledgments for the roles played by different actors. Communicate that information to future users.

In this sense, good identifier management practices become a core aspect of professional conduct not only for data creators, but also for data reusers. In data aggregation and reuse scenarios, it is of critical importance to maintain contextual integrity of source datasets. Tracing provenance and other aspects of contextual integrity can improve the credibility and transparency of "big data" studies that amalgamate multiple smaller archaeological datasets (1).

Maintenance of contextual integrity can also encourage greater social and professional recognition for the collaborative processes behind data creation. The CARE (Collective benefit, Authority to control, Responsibility, Ethics) data principles, together with related efforts such as "traditional knowledge" labels, aim to formalize aspects of ethical research practice, especially in the context of benefits sharing, and recognition of indigenous peoples (42). In this regard, identifiers in these source datasets may document authorship and other aspects of professional, social, and cultural recognition. Identifiers in these source datasets may also describe intellectual property licenses and expectations for reciprocity and recognition for members of indigenous and other descendent communities. Data reusers therefore have a responsibility to adequately curate identifiers in the datasets they reuse. That curation helps document both the analytic and the social context of source data and would encourage continued recognition of roles and responsibilities in all aspects of data creation, curation, and reuse. This need will be especially acute for domain specialists, since as discussed, their research interests often span across regions and involve inputs from multiple field projects, each of which may have their complex web of expectations that need to be respected.

Conclusions

Data quality expectations shift and evolve as the universe of intended uses expands and becomes more open ended. Practices that adequately serve small teams of researchers may not adequately meet the needs of a wider community. Improving archaeological data practices so that they can support synthetic and integrative research requires greater attention to identifier practices. Identifiers make it possible to reliably and precisely relate information in a given dataset to information in other datasets. Identifiers therefore play a key role in shaping the contextual integrity and quality of research data. As the interpretive and analytic value of archaeological data heavily depends upon contextual integrity, archaeologists need to adopt less siloed and more collaborative orientations in the creation and curation of data.

Data Availability. There are no data underlying this work.

1 J. Huggett, Is big digital data different? Towards a new archaeological paradigm. *J. Field Archaeol.* **45** (suppl. 1), S8–S17 (2020).
2 K. W. Kintigh, The promise and challenge of archaeological data integration. *Am. Antiq.* **71**, 567–578 (2006).
3 I. M. Faniel *et al.*, Identifying opportunities for collective curation during archaeological excavations. *Int. J. Digit. Curation* **16**, 17 (2021).
4 E. Yakel, I. M. Faniel, Z. J. Maiorana, Virtuous and vicious circles in the data life-cycle. *Inf. Res.* **24**, 821 (2019).
5 N. Smale, K. Unsworth, G. Denyer, E. Magatova, D. Barr, A review of the history, advocacy and efficacy of data management plans. *Int. J. Digit. Curation* **15**, 1–29 (2020).
6 I. M. Faniel *et al.*, Beyond the archive: Bridging data creation and reuse in archaeology. *Adv. Archaeol. Pract.* **6**, 105–116 (2018).
7 I. M. Faniel, E. Yakel, "Practices do not make perfect: Disciplinary data sharing and reuse practices and their implications for repository data curation" in *Curating Research Data, Volume 1: Practical Strategies for Your Digital Repository*, L. R. Johnston, Ed. (Association of College and Research Libraries, Chicago, IL, 2017), pp. 103–125.
8 L. Atici, S. W. Kansa, J. Lev-Tov, E. C. Kansa, Other people's data: A demonstration of the imperative of publishing primary data. *J. Archaeol. Method Theory* **20**, 663–681 (2013).
9 E. C. Kansa, S. W. Kansa, B. Arbuckle, Publishing and pushing: Mixing models for communicating research data in archaeology. *Int. J. Digit. Curation* **9**, 57–70 (2014).
10 S. W. Kansa, L. Atici, E. C. Kansa, R. H. Meadow, Archaeological analysis in the information age: Guidelines for maximizing the reach, comprehensiveness, and longevity of data. *Adv. Archaeol. Pract.* **8**, 40–52 (2020).
11 G. Gattiglia, Think big about data: Archaeology and the Big Data challenge. *Archäologische Informationen* **38**, 113–124 (2015).
12 M. D. McCoy, Geospatial Big Data and archaeology: Prospects and problems too great to ignore. *J. Archaeol. Sci.* **84**, 74–94 (2017).
13 M. D. McCoy, T. N. Ladefoged, New developments in the use of spatial technology in archaeology. *J. Archaeol. Res.* **17**, 263–295 (2009).
14 R. H. Wilshusen, M. Heilen, W. Catts, K. de Dufour, B. Jones, Archaeological survey data quality, durability, and use in the United States: Findings and recommendations. *Adv. Archaeol. Pract.* **4**, 106–117 (2016).
15 S. Wolverton, Data quality in zooarchaeological faunal identification. *J. Archaeol. Method Theory* **20**, 381–396 (2013).
16 L. A. Lodwick, Agendas for archaeobotany in the 21st century: Data, dissemination and new directions. *Internet Archaeol.*, https://doi.org/10.11141/ia.53.7 (2019).
17 R. Y. Wang, V. C. Storey, C. P. Firth, A framework for analysis of data quality research. *IEEE Trans. Knowl. Data Eng.* **7**, 623–640 (1995).
18 R. Y. Wang, D. M. Strong, Beyond accuracy: What data quality means to data consumers. *J. Manage. Inf. Syst.* **12**, 5–33 (1996).
19 L. Cai, Y. Zhu, The challenges of data quality and data quality assessment in the Big Data era. *Data Sci. J.* **14**, 2 (2015).

20 S. Sadiq, M. Indulska, Open data: Quality over quantity. *Int. J. Inf. Manage.* **37**, 150–154 (2017).
21 B. Marwick *et al.*, Open science in archaeology. *SAA Archaeol. Rec.* **17**, 8–14 (2017).
22 B. Marwick, Computational reproducibility in archaeological research: Basic principles and a case study of their implementation. *J. Archaeol. Method Theory* **24**, 424–450 (2017).
23 L. Brenskelle *et al.*, Identifiers as mechanisms for linking archaeological data across repositories. *Biodivers. Inf. Sci. Stand.* **2**, e26471 (2018).
24 H. Lau, S. W. Kansa, Zooarchaeology in the era of big data: Contending with interanalyst variation and best practices for contextualizing data for informed reuse. *J. Archaeol. Sci.* **95**, 33–39 (2018).
25 E. Plomp, Going digital: Persistent identifiers for research samples, resources and instruments. *Data Sci. J.* **19**, 46 (2020).
26 N. Juty *et al.*, Unique, persistent, resolvable: Identifiers as the foundation of FAIR. *Data Intelligence* **2**, 30–39 (2020).
27 D. Carruthers, Data from "Pınarbaşı 1994: Animal Bones". Open Context. Available at https://n2t.net/ark:/28722/k2zs2vk0z. Deposited 25 March 2006.
28 N. Russell, Data from "Pınarbaşı Bird Remains". Open Context. Available at https://doi.org/10.6078/M71R6NMG. Deposited 25 August 2020.
29 M. J. LeFebvre *et al.*, ZooArchNet: Connecting zooarchaeological specimens to the biodiversity and archaeology data networks. *PLoS One* **14**, e0215369 (2019).
30 B. Marwick, S. E. P. Birch, A standard for the scholarly citation of archaeological data as an incentive to data sharing. *Adv. Archaeol. Pract.* **6**, 125–143 (2018).
31 K. W. Kintigh *et al.*, Grand challenges for archaeology. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 879–880 (2014).
32 T. Berners-Lee, J. Hendler, Publishing on the semantic web. *Nature* **410**, 1023–1024 (2001).
33 C. Shapiro, H. R. Varian, *Information Rules: A Strategic Guide to the Network Economy* (Harvard Business School Press, 1998).
34 J. Hendler, J. Golbeck, Metcalfe's law, web 2.0, and the Semantic web. *J. Web Semant.* **6**, 14–20 (2008).
35 A. E. Thessen *et al.*, From reductionism to reintegration: Solving society's most pressing problems requires building bridges between data types across the life sciences. *PLoS Biol.* **19**, e3001129 (2021).
36 E. A. Schultes, G. O. Strawn, B. Mons, "Ready, set, GO FAIR: Accelerating convergence to an internet of FAIR data and services" in *Proceedings of the XX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2018)* (Moscow, Russia, 2018), pp. 19–23. http://ceur-ws.org/Vol-2277/paper07.pdf.
37 N. Davies *et al.*, Internet of samples (iSamples): Toward an interdisciplinary cyberinfrastructure for material samples. *Gigascience* **10**, giab028 (2021).
38 J. C. Wallis, E. Rolando, C. L. Borgman, If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS One* **8**, e67332 (2013).
39 B. S. Arbuckle *et al.*, Data sharing reveals complexity in the westward spread of domestic animals across Neolithic Turkey. *PLoS One* **9**, e99845 (2014).
40 M. D. Wilkinson *et al.*, The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
41 T. M. Christian, A. Gooch, T. Vision, E. Hull, Journal data policies: Exploring how the understanding of editors and authors corresponds to the policies themselves. *PLoS One* **15**, e0230281 (2020).
42 S. R. Carroll, E. Herczog, M. Hudson, K. Russell, S. Stall, Operationalizing the CARE and FAIR principles for indigenous data futures. *Sci. Data* **8**, 108 (2021).