Primary Sources as Linked Data: Exploring Motives Across the Sciences and Social Sciences

Marsh, Diana E. University of Maryland, USA | dmarsh@umd.edu
Fenlon, Katrina University of Maryland, USA | kfenlon@umd.edu
Sorensen, Amanda H. University of Maryland, USA | asorens1@umd.edu
Wise, Nikki M. University of Maryland, USA | nwise@umd.edu

ABSTRACT
While long recognized in the humanities, there is growing recognition in the sciences and social sciences that primary sources—as diverse as manuscripts, photographs, cultural belongings, and specimens—hold vast data about scientific and human knowledge for use in scholarship, community research, and global knowledge. Yet, data embedded in these sources are largely disconnected from the systems of discovery, access, and structured data that support reuse and insights across globally dispersed repositories. In this paper, we share select findings of a systematic review to explore the use of primary sources, and the data embedded in them, via linked data across the sciences and social sciences. Our results confirm the use of a variety of primary source data across diverse disciplines, particularly those requiring longitudinal studies and data integration from diverse repositories and contexts. We highlight how linked data are understood to: connect collections to communities; support highly granular credit, attribution, and assessment of impact; and interrelate diverse sources of knowledge. While these results suggest the value of linked data for the specific research needs of anthropology, the effectiveness of linked data in achieving these objectives and the suitability of this approach for a diversity of institutions and communities need further study.

KEYWORDS
Data reuse; archives; primary sources; science; social science

INTRODUCTION
While long recognized in the humanities, there is now growing recognition in the sciences and social sciences that primary sources—as diverse as manuscripts, photographs, cultural belongings, and specimens—hold vast data about scientific and human knowledge for use in contemporary scholarship, community research, and global knowledge. Yet, data embedded in these sources are largely disconnected from the systems of discovery, access, and structured data platforms that support reuse and new insights across globally dispersed repositories. Linked data—a constellation of standards and tools for representing data in such a way that they can meaningfully and actionably connect to data and knowledge across the broader Web (van Hooland & Verborgh, 2014)—have been widely discussed as a promising option for representing data from and about primary sources. However, the prevalence of linked data implementations and their effectiveness is understudied, particularly as we look beyond the humanities and cultural heritage toward implementations that connect data and knowledge more broadly across disciplines.

In this paper, we share select findings of a systematic review of linked data initiatives that are beginning to exploit these potentialities and lay the groundwork for field-wide approaches. We ask: how are scientists and social scientists using primary sources, and especially linked data derived from primary sources?

This work began as part of the first year of an NSF-funded project (NSF CA-SR award #2314762) called *Linking Anthropology's Data and Archives* (LADA), which attends to this problem within the discipline of anthropology. Through the three-year project, we aim to improve 1) how anthropologists create and curate their own sources to support reuse; 2) how professional domains of data curation and archives work with analogue anthropological sources, convert sources to linked data, and connect data to broader knowledge ecosystems; and 3) how to adapt data curation and reuse best practices from other disciplines to anthropology's unique needs and considerations.

Of course, facilitating *anthropological* data reuse from analogue sources will advance *anthropological* research across that domain, but in our first year of the work we began to understand anthropology's potential as a central case study for cross-disciplinary understandings of this problem. Anthropology, particularly in its Americanist traditions, encompasses cultural anthropology and linguistic anthropology–which span humanistic and social science approaches–and archaeology and biological anthropology–which span social science and science approaches. Anthropology also notoriously grapples with a long colonial history and attendant range of ethical concerns, while also driving many decolonizing and collaborative movements now shaping other disciplinary domains. It therefore serves as a useful central node in which to explore the untapped potential of previously collected records, and of data derived from them, to support novel lines or reuse both in spheres of interdisciplinary research and in

community spaces. Through our project, we aim to both learn from, and speak to, a wide range of analog collections-based linked data work in the sciences and social sciences.

The first phase of this project therefore constitutes a systematic review of linked data projects across those domains. We reviewed an initial sample of 303 projects, winnowing this sample to a final set of 35 projects, which we subjected to qualitative coding to understand how and why they are deploying linked data in service to research in the sciences and social sciences. Below we describe facets of scoping for this study, including our definitions of primary sources. Our results confirm the use of a wide variety of primary sources and data derived from them across diverse scientific and social scientific disciplines, particularly those tackling research problems that demand longitudinal study and the integration of data from a wide range of sources and disciplinary contexts. However, the use of linked data for this purpose remains more exploratory than widespread. The ideals that motivate participation in linked data among scientific and social scientific projects are numerous, and include facets of data curation, discovery and access, interpretation and analysis, and data reuse. These align with the discourse around linked data in cultural heritage and digital humanities. We highlight a few themes that stand out among the objectives of linked data projects in the sciences and social sciences, including how linked data are understood to connect collections to communities; how linked data support highly granular credit, attribution, and assessment of scientific impact; and how linked data are understood to interrelate diverse sources of knowledge, including primary and secondary data and literature. While these results suggest the value of linked data for the specific research needs of anthropology, the effectiveness of linked data in achieving these objectives and the suitability of this approach for a diversity of institutions and communities need further study.

PRIOR WORK

The use of linked data to represent primary sources held in libraries, archives, and museums (LAMs) has been a growing topic of discourse for more than a decade in cultural heritage and library and information sciences research and practice, particularly at the intersection with digital humanities scholarship. Various review studies have charted the uptake of linked data in LAMs, including both in support of regular operations (e.g., the infusion of linked data into bibliographic description, technical services, and collections management systems) and for research-oriented initiatives to explore the potential of linked data to serve innovative modes of access to and use of LAM collections (Hyvönen 2012; Hawkins, 2022; Bikakis et al., 2021; Gaitanou et al., 2024). Numerous initiatives in digital humanities scholarship have demonstrated the applications of linked data to representing historical and humanistic knowledge, along with strengths and weaknesses of the approach (Nurmikko-Fuller, 2024).

With a few exceptions, empirical studies of linked data practices and their impacts that look across disciplinary boundaries are lacking. Hawkins (2022), for example, notes that efforts toward increasing research data reuse in the sciences, particularly the FAIR guidelines (Wilkinson et al., 2016), have helped spur the rise of linked data in LAMs more broadly. This paper sits at an understudied intersection among otherwise divergent disciplinary practices: the application of linked data to the representation of primary sources, but within the sciences and social sciences rather than in the humanities or cultural heritage. To frame this study, we examine three related areas of work: broad movements in open science and data reuse that contextualize the specific focus of this study; the use of primary sources and data derived from them as evidence in the sciences and social sciences; and, more specifically, prior research on the deployment of linked open data in these domains.

Open data and reuse across disciplines

This study of how primary sources serve as data within the sciences and social sciences emerges at the confluence of two streams of widespread, decades-long effort to increase open access in memory and research institutions. Libraries, archives, museums, and related research disciplines have widely worked to increase the possibilities of digital access to collections and cultural data, not only for public access to cultural heritage and historical knowledge, but also to support computational approaches to research, particularly in the humanities. At the same time, the movement toward open access to knowledge took root within the realms of scholarly and scientific publishing, with the goals of increasing public understanding of and trust in science, ensuring research reproducibility, and growing opportunities for novel data reuse. These movements have produced a wide range of systemic supports for the use of data across disciplines, including but not limited to supportive policies (Nelson, 2022); high-impact guidance for the production and use of data, such as the FAIR (Wilkinson et al., 2016) and CARE principles (Carroll et al., 2020); a growing population of credentialed professionals in a wide range of institutions with expertise in domain-specific approaches to preserving and making data accessible (Kouper, 2016); and a growing host of repositories for maintaining access to data over time.

The Collections as Data initiative (2016-2023) explored convergences among diverse, international efforts to make memory institution collections accessible as data (Padilla et al., 2023, 2). While this and many other strands of research in data curation evince significant progress toward ethically grounded, useful, sustainable approaches to computationally amenable cultural collections, they also highlight persistent, unmet needs among all stakeholders.

For researchers and practitioners there remains a profound gap between the extent of resources held in cultural institutions and the extent available digitally, only a small fraction of which are also available as computationally amenable data. Across fields, there is a pervasive need to lower technical barriers, and for training, skill-sharing, and communities of practice to support the development and use of digital collections and data (Dallas 2016; Jaillant, 2022; Mayernik, 2016). Various programs and services have emerged to meet these needs, including the cross-institutional Data Curation Network (Johnston et al., 2018) and an array of conferences and communication venues to connect researchers and practitioners into networks of practice, such as LD4L (Gaitanou et al., 2024). But these are still relatively small communities, and reflect efforts that are often more experimental than operationally rooted in LAMs. Despite growing attention to community-centered approaches to collecting and data sovereignty, there also remain many unanswered questions about how best to engage communities in the development and representation of data collections that pertain to them, and about how to address and contextualize collections from which data derive effectively and ethically (Walter et al., 2019). Further, scholarship has recognized that descriptive practices amongst collections and repository types vary widely. For example, archival description methods provide a large degree of flexibility and "accommodates lower quality descriptive records" (Weidman, 2023).

In addition to these theoretical and pragmatic gaps, there is a need for a broader, empirically grounded view of how people in different contexts and across disciplinary traditions do or want to use primary source collections as data.

Primary sources and data reuse

The use of primary sources as evidence is much more common in the humanities, but a growing body of work acknowledges the value of primary sources and their embedded data to support contemporary scientific and social scientific research (Sorensen et al., 2023; Kelly et al., 2022). Scientists and scholars across disciplines rely on a diversity of types of evidence to support argumentation and knowledge production. Data extracted from primary sources are being used to re-examine or refine interpretations, facilitate comparison, and/or validate conclusions. Yet, despite increasing both critical and applied attention to the digital 'datafication' of primary source data (Flensburg & Lomborg 2023; Yang 2024), most primary sources and their data are often disconnected from the web, and the digital systems of discovery and access that support research and community work. Many primary sources are physical archival documents and museum collections held in institutions across the globe whose collections are not yet digitized. Despite great strides in digitization technologies, a fraction of the world's collections are imaged and available online. For instance, the National Archives and Records Administration as of the time of writing has only 2% of its collections digitized, which comprises nearly 250 million scans. Despite vast strides in mass digitization workflows and conveyor scanning, the Global Collection Group found that of 73 museums and 1,147,934,687 specimens surveyed, only 16% of collections had digitally discoverable records (Johnson et al., 2023; Sorensen, 2023). Second, primary sources are represented and described in a wide range of standards and content management systems that often do not align.

Research on 'data recovery' and 'data rescue' has shown that information professionals and researchers are mobilized across many domains to recover useful data from analog and digital primary source collections (Sorensen et al., 2023; Kelly et al., 2022). Research on crowdsourcing focused specifically on historical scientific data has shown the potential of scientific data such as climatic and weather data to be extracted from historical records and reused in contemporary research and discourse (Brunet & Jones, 2011; Wippich, 2012). Research on scientific data sharing and museum knowledge infrastructures has likewise shown the growing interest in primary source data reuse across a range of natural history fields (Thomer, 2022). Previous work has indicated that challenges in data reuse will benefit from archival perspectives, data rescue and lifecycle thinking, greater dialogue across scientific disciplines, and community-driven approaches (Sorensen et al., 2023). In anthropology and other collecting fields in natural history, scholars have argued that amidst colonial collecting histories and the sloughs of backlogged materials in repositories, archival and data reuse, as opposed to new collecting, represents an important ethical stance (Buchanan 2019; Kirakosian & Bauer-Clapp 2017; Cliggett, 2013). There has been some information scholarship in this vein on the reuse of archaeological data in the form of museum collections (Daniels, 2014; Kriesberg et al. 2013), the development and challenges of linked data documenting archaeological materials (Schmidt et al., 2022), and survey research has revealed that science-based anthropology fields are increasingly drawing on analog archives (Marsh et al., 2023).

We circumscribe our focus on a wide range of *primary sources*, or 'things' (Henare et al., 2007) that can provide direct and immediate evidence—whether historical or contemporary—of phenomena under study. Data are derived from primary sources by interpretive, subjective processes, which may often involve different aspects of original sources: of their form and content, and frequently of interpretive context that get transcribed into the data as well (Furner, 2004). The "primariness" of a source is a matter of degree (Paulus Jr., 2012). Specifically, we scoped our study to include original documentary, artifactual, or specimen sources of evidence held in memory institutions including libraries, archives, museums, and repositories. While primary sources are often maintained and provided

in other contexts, constraining our study to evidence held in memory institutions limits our view to items being managed and treated as evidence in domains most relevant to the overarching study.

Linked Data in the Sciences and Social Sciences

The concept of linked data encompasses a range of standards, tools, and activities geared towards interconnecting datasets across the Web. In the linked data or semantic web paradigm, distributed data are connected in such a way that their semantics and original contexts are maintained, yet they are interoperable and amenable to integrated computational analysis. For this study, we focus on a narrow conception of linked data as those that conform to Linked Data Principles, and which obtain a level of 4 or 5 stars in the 5-Star Linked Open Data scheme originally offered by Tim Berners-Lee in 2006 (W3C, n.d.; Berners-Lee, 2006). By this definition, data must be published on the Web using deferenceable HTTP URIs to name entities, express data using open Web standards promulgated by the W3C (RDF, SPARQL, JSON-LD, etc.), and ideally include URI links to other related data and entities.

The decision to focus on linked data as one among various alternative modes of representation for digital primary sources and data stems from the growing interest in and discourse around linked data in cultural heritage institutions, where anthropological evidence often resides. This discourse in turn has long been spurred by the alignment of linked data with the high-level objectives of cultural institutions—in the open accessibility of knowledge, the preservation of provenance and context, and in making connections across collections (Davis & Heravi, 2021; van Hooland & Verborgh, 2014). Yet, in the decade plus since the term began gaining currency in libraries, archives, and museums, the movement toward linked data has confronted numerous obstacles (Davis & Heravi, 2021; Kansa, 2015; Geser, 2016). Linked data is one among many alternative paths forward for anthropological data representation; and this study is a first step in evaluating the prevalence and use of linked data in relevant domains of research. But this is not a commitment to linked data as the only or best option for data representation.

While there have been systematic, empirical, and other intensive studies of the use of linked data in the humanities and cultural heritage, applications in the sciences and social sciences are relatively understudied. A key exception is Da Silva (2018), which looked at impacts of linked data on social science and humanities methodologies. Yet, there are signs of significant and widespread activity that motivate this research. Linked data projects are taking place in fields as diverse as biology, chemistry, and health sciences, as a means to ask broad questions of field data from otherwise disparate, isolated sources (for instance PubChem (n.d.) or PubMed (n.d.)). Examples of the biology and chemistry sites include the European Biology Institute (EBI) Search RESTful Web Services API (EBI, n.d.), the ChEMBL database's Data Web Services (Willighagen et al., 2013), and Bio2RDF (Bio2RDF v2.7a., n.d.; Callahan et al., 2013). In health sciences, The Cancer Genome Atlas (TCGA) Computational Tools (Saleem et al., 2014), BioFed query engine (Hasnain et al., 2017), and the Linking Open Drug Data (LODD) project (Jentzsch et al., 2009; Samwald et al., 2011) all make biomedical and pharmacological data available through linked data infrastructures. In the social sciences there has likewise been movement toward linked data. In archaeology, the ARIADNE Linked Data Cloud, the Archaeology Data Service (ADS), and Open Context all suggest the recognized value of the approach (Geser, 2016; Tudhope et al., n.d.; Open Context, n.d.). Linked data deployment in the social sciences more broadly is less visible. With some exceptions (such as Hajra & Tochtermann, 2017 and Zapilko et al., 2016), most prior research on linked data applications takes the form of domain-specific case studies, seldom looking across disciplinary divides.

We included both projects that provide linked data, and those which seem to leverage linked data in some capacity, sometimes on the back end of their infrastructures and in ways that are not immediately visible to users. Both types of projects participate in the linked data ecosystem and help illuminate the opportunities for primary sources as linked data. Because we did not systematically assess the open licensing status of projects and data under consideration, we refer to the more general concept of *linked data* (rather than *linked open data*) throughout this study. The majority of projects we reviewed provide data freely and openly under different licensing and access conditions; we excluded projects providing data that were not freely accessible to public use.

In practice it frequently proved difficult to determine the extent to which a project offers or relies upon linked data. Many initiatives in different domains explicitly or implicitly do linked data work, aiming to do large-scale data integration, but not always relying on W3C standards. For example, numerous interrelated initiatives in the Earth Sciences and biodiversity have long recognized the value and importance of interconnecting data from different sources and different research disciplines into a more complete picture of ecosystems over time. Efforts to aggregate and integrate data in these domains are multifaceted and multilayered, and these tend to intersect with the linked data universe in various, sometimes obscure ways. There are widely used standards in these domains, such as DarwinCore (n.d.), which are capable of expression in linked data standards such as RDF, but which are not necessarily implemented as linked data in their most common use cases. There are many data providers and data aggregators that deploy APIs (in line with linked data principles) but rely predominantly on domain-specific rather than Web standards for data representation.

Indeed, much of the scientific work being undertaken with aggregated and linked data is being led by informatics professionals and scientists in the world's natural history museums. As a recent article in *Science* outlined, museums across the globe are being urged to integrate their collections with the networks and platforms, such as the Global Biodiversity Information Facility (GBIF, n.d.) Atlas of Living Australia (n.d.), Integrated Digitized Biocollections (iDigBio, n.d.), the Distributed System of Scientific Collections (n.d.), and Biodiversity Collections Network (BCoN) (n.d.; Johnson et al., 2023). This data in aggregate reveals crucial baseline and temporal data about climate, species health, and global biodiversity. Much of this work integrates unified standards such as DublinCore, but does not qualify as linked data.

We debated heavily whether to include aggregators in this vein such as iDigBio (n.d), GBIF (n.d) and GeoCASe (n.d.), and made different decisions in each case, based on how readily apparent the linked data aspect is. Such aggregators can or do participate in broader linked data networks in various ways, often through experimental or one-off initiatives. Where relevant, we included these more specific initiatives in our dataset. Regardless of whether each initiative was ultimately included in the dataset, we draw inspiration from these and other intersecting worlds of data integration and will prioritize them for future study as part of our overarching research project.

METHODS

The objective of this research was to understand the use of primary sources, and the data embedded in them, via linked data across the sciences and social sciences. We conducted a systematic search and review process that was also informed by systematic review and search criteria methods (Grant & Booth, 2009; Granikov et al., 2021). This method was selected due to its strength in identifying "what is known," the current limitations of the unit of analysis, and "recommendations for practice," all of which align with the impetus of our year work as described within our associated grant application (NSF CA-SR award #2314762; Grant & Booth, 2009). First, the team formulated our guiding question for the search and review: How is linked data being used to share primary sources representations and data derived from them in the sciences and social sciences? We established each linked data project as our unit of analysis and defined our search criteria (see Table 1):

	Inclusion Criteria	**Exclusion Criteria**
Type of project	LD project drawing on primary sources, where a primary source is an original, documentary, artifactual, or specimen source of evidence held in memory institutions such as LAMs and repositories	Not using LD; not drawing on primary sources
Population	Organizations, institutions, team that created, developed, maintain the projects (leaders)	Participant organizations, institutions, teams, process who did not originate (joiners)
Setting	Connects to broader Web: Analog or digitized materials (not born-digital) represented as linked data online; Primary source materials have to be held in a memory institution repository	Not openly available online: analog or intranet infrastructure; Primary sources located outside of memory institutions and repositories
Intervention	Sharing data within the above fields through LD; data reuse in the above fields through LD	Sharing data through other means (i.e. analog, databases without linked data, etc.)
Types of publication	Websites, journal articles, informal publications, funding reports	n/a
Language	English	Other languages
Date	2018 onward	Published prior to 2018

Table 1. Inclusion and Exclusion Criteria

The team then identified potential information sources for both potential projects and discipline lists. These included drawing on information generated from the Recovery and Reuse of Archival Data for Science (RRAD-S) project, which was conducted in collaboration with the National Agriculture Library. We also incorporated recommendations from the Council for the Preservation of Anthropological Records Working Group (who are serving in an informal advising capacity). Then, we conducted a systematized search and review using Google, Google Scholar, and later Bing (due to seeming bias in search results by researcher). We further utilized snowball sampling to complete the initial dataset based on sources reviewed from all of the above techniques. Any and all

documentation describing the linked data projects, as our unit of analysis, were understood as under the purview of our analysis, and we examined sources until we reached saturation pertaining to the codes we applied.

The disciplines were restricted to science and social sciences. The original list of social science disciplines that were searched came from the disciplines listed in *An Assessment of Research-doctorate Programs in the United States--Social and Behavioral Sciences* (Jones & Coggeshall, 1982). These were psychology, sociology, geography, economics, anthropology, political science. The social science discipline was expanded to include disciplines identified by the National Center for Education Statistics (n.d.), so that data about these fields could be compared across other studies. Eleven social sciences disciplines were searched in total. The sciences discipline list was created using a combination of disciplines from our prior RRAD-S project and from Kelley et al. (2022). We searched 25 science disciplines and 11 social science disciplines in total, acknowledging that the edges of these areas are often blurry. In the sciences we reviewed relevant projects in agriculture, astronomy, biodiversity, biogeography, biology, botany, chemistry, climatology, earth sciences, ecology, epidemiology, environmental sciences, forestry, geography, geology, geoscience, hydrology, marine science, meteorology, natural history, oceanography, ornithology, physics, water research, zoology. In the social sciences we reviewed efforts in anthropology, archaeology, criminology, demography, economics, geography, international relations, political science, psychology, sociology, and urban studies.

Each investigator was assigned a different list of disciplines to search for. A sample search query in Google Scholar, Google, and Bing is as follows: "Linked data" AND "archiv" AND "anthropology." Each project was analyzed for inclusion/exclusion criteria by its entry coder. That data set was then expanded via snowball sampling: as we reviewed each case and its associated web presence and documentation for our selection criteria, we occasionally identified additional projects not included in our sample. The team erred on the side of including projects that appeared to meet the criteria or self-identified as meeting our criteria, to look broadly at possible cases. For instance, many projects (particularly in some domains such as criminology) self-identify as using "linked data" but were in essence drawing on database structures, aggregator tools, or interoperable data sharing, that did not draw upon linked open data as we defined it. We decided to focus on projects that were active in the past five years (between 2018 and the present) in order to analyze recent work, acknowledging that such a cutoff is somewhat arbitrary. While the concept of linked open data was developed in the mid-2000s (Berners-Lee et al., 2001) and discourse around the concept surged in the GLAM fields in the 2010s (Davis & Haravi 2021, 5), over the past five years the domain has reached a state of maturity in which implementation is catching up to discourse. As we discuss later, we may wish to include a wider range in future stages as we explore the longevity and sustainability of these efforts.

After that data collection process, our full sample included 303 projects. Coding drew on qualitative content analysis (Zhang & Wildemuth, 2016), and included both predetermined codes based on initial research, and iterative inductive coding during deeper analysis of open-ended project descriptions. The team then built out an initial codebook as a group to include core areas of interest such as types of primary sources, primary discipline, and the purpose of linked data. Consensus was then built about tightening inclusion/exclusion criteria and the meanings of codes. Each project was coded for both inclusion/exclusion and information entered for our preliminary set of additional coding categories by a first entry coder. Coding and the conformance to inclusion/exclusion criteria was evaluated for quality and completeness by at least two coders, including one PI. When there was disagreement about a project's eligibility or coding, the team met to discuss edge cases to build consensus and understanding about the meaning and application of the criteria and the codes. After this iterative coding process, the list of projects that met our inclusion/exclusion criteria was winnowed to a final sample of 34. Those were annotated and described in more detail by each cases' entry coder, and all entries were collectively discussed by the whole team. Finally, each PI conducted deeper inductive coding of core coding categories (disciplines and primary source types by the PI and linked data purpose by the Co-PI) to yield the preliminary findings shared here.

FINDINGS

Disciplines and Primary Sources

In total, we characterized 17 of these projects as primarily emanating from or serving 'social science' disciplines, 16 from 'science,' and one, Europeana, as weighted equally to both. We observed a huge range of primary sources being leveraged in linked data spaces across our 34 cases. Most commonly, we categorized these source collections as **paper archival records** (which included analog datasets, manuscripts, fieldnotes), **specimens** (which included both scientific collections and samples such as cores or samples derived from them), **objects/belongings** (which included human-produced items or artifacts, **photographs** (including analog negatives, prints, aerial images), and **historical or rare book publications** (including historical monographs and rare print materials as well as reports). A few additional categories included murals or wall paintings (alluding to architectural or site-based works, and audio recordings. We kept disciplinary categorizations of (usually museum-held) objects as human-made belongings or specimens from the natural world to attend to the ethical and dehumanizing potential for doing otherwise, but it is

important to note that "objects" of various kinds were by far the most commonly cited in projects across domains. Photographs, while less often cited, were equally utilized by projects in the sciences and social sciences.

There is some disciplinary difference among the citation of these collections as primary source data– for instance more use of paper and archival materials, as well as historical publications, among social science projects. The projects in the sciences tended to rely more heavily on specimens overall (Table 2).

	Sciences	**Social Sciences**	*Total occurrences*
Paper/Archival	7	13	19
Specimens	16	2	18
Objects/Belongings	3	9	12
Photographs	4	4	8
Historic Publications	2	5	7

Table 2. Most Common Primary Sources by Disciplinary Category

Specimens as diverse as herbaria specimens, faunal remains, fossils, corals, pollen, sediments, ice cores, and tree rings were noted as core collections for linked data projects in biodiversity, biology, ecology, biological anthropology, ocean sciences, paleobiology, earth science, geology, entomology, and botany. Human-produced objects and belongings were cited among a few of the scientific projects relating to biodiversity, biological anthropology, and zooarchaeology (necessarily human-interested fields). In the social sciences, projects in archaeology, cultural anthropology, geography, and linguistics noted the use of human-made belongings and objects. Paper and archival records were used by projects in disciplines such as archaeology, botany, earth sciences and demography in the social sciences, and in disciplines as diverse as botany ecology, entomology, evolutionary biology, geology, and zoology. Photographs were used in projects serving a range of biology and biodiversity fields in the sciences, and archaeology and anthropology fields in the social sciences.

Why Linked Data?

Within our dataset, we coded for project-based descriptions about why a given project chose to use linked data. We looked at "about" pages on project websites and associated scholarly publications detailing the technical specifications and their rationale.

Of course, a number of projects noted the technical capabilities of linked data in a general sense. A number of the dataset's projects cite core qualities, such as promoting access to data, trackability, and interoperability. For example, the Biodiversity Heritage Library (BHL) "works to bring historical literature into the modern network of scholarly research by retroactively assigning DOIs (digital object identifiers) and making this historical content more discoverable and trackable" (Kalfatovic et al., 2023). The International Digital Dura-Europos Archive (IDEA) "aims to virtually reassemble and improve the accessibility of digital archival content related to the archaeological site of Dura-Europos" (IDEA, n.d.). This access, trackability, and interoperability is intended to "overcome technological silos" (Cornut et al., 2023), and to "automate the handling of complex data, bring together information gathered from various sources and enable easy access to 'isolated' information" (Nundloll et al., 2022).

But upon deeper analysis, four more complex core themes emerged as core aspirations for linked data across our dataset: Data Curation; Discovery and Access; Reuse; and Interpretation and Analysis. Importantly, the core affordances we coded within each of the latter three themes are derived from capabilities described within the Data Curation theme. In other words, the core foundations of data curation not only make possible many of the other qualities of linked data work, but when done well can generate more effective and impactful work in other domains. Within these four themes, three surprising characteristics of linked data stood out across our cases as distinctive: to build connection with communities, to give credit and attribute data, and to interrelated sources across data formats, spanning historic publications, archival documents, specimens, and museum collections. Figure 1 below overviews themes from the reasoning that we gathered across the dataset.

Figure 1. Objectives of Implementing Linked Data Coded by Purpose

Connection to Communities (Discovery & Access)

A handful of projects within the dataset articulate their goals in relation to communities. For instance, PARADISEC aims to connect research with the communities they have worked in, but also to connect communities with their data in appropriate formats (Arkisto, n.d.). Similarly, projects discuss the "democratization" of data. DigIn, for example, aims to "contribute to democratization of biodiversity data and the diversity of people with access to it" (DigIn, n.d.) Europeana is also interested in the public's novel uses of linked data: "We love to see the creative ways you use what you find here, whether it's teachers developing resources for lessons, developers using our open-source API to make games, or culture lovers creating gifs and telling stories" (Europeana, n.d.). As disciplinary approaches beyond anthropology increasingly draw on relational and collaborative practices, linked data's potential to share data across communities is noteworthy.

Credit and Attribution (Data Curation)

Projects within the dataset also describe how linked data can support systems of highly granular credit and attribution, allowing for the attribution of specific data points, contextual insights, data manipulations, and other aspects of provenance to specific researchers. For example, Bionomia discusses this point, stating that "If sufficient numbers of people claim their specimens and attribute them to others, we may unwittingly develop reconciled authority files to help museum staff employ an integrated approach that incorporates look-ups of disambiguated

people names" (Bionomia, n.d.). While it is established that credit and attribution are important for data reuse, these aspects are less frequently associated with the value of linked data in humanities and cultural heritage contexts.

Interrelating Sources and Context (Interpretation and Analysis & Reuse)

Projects within our dataset aim to interrelate various types of sources and their context. While predictable among historians, we noted this across a range of social science and science examples. For instance, the BHL draws together fieldbooks, biodiversity literature, photographs, illustrations, and manuscripts (Biodiversity Heritage Library, n.d.). The Archaeology Data Service interlinks field reports, datasets, GIS data, and geophysical observations (Binding et al., 2022). Open Context also draws on a variety of data format types, including bound paper, faunal collections, photographs, and objects/artifacts (Open Context, n.d.). Attempts to link and interrelate across a variety of source types speaks to the silos that typically exist between libraries, archives, and museums based on material type. Researchers and practitioners are drawing on linked data to support work across these silos.

DISCUSSION

In addition to the technical affordances of linked data, the projects in our dataset articulate bigger picture 'whys.' But we know less here about the true sociotechnical and disciplinary cultures in which this work is taking place, beyond our own personal backgrounds in these fields. We have some initial hypotheses about some of these differences and convergences that we hope to explore in the next phase of work. For instance, we have a preliminary sense that a lot of the big picture logic for undertaking linked data work across both the sciences and social sciences is fundamentally human-centered. Of course, that goes without saying in the social sciences, which are, by definition, about human worlds and impacts.

But in the sciences, this is perhaps interesting. Much of the reasoning behind larger primary source-driven linked data efforts in the sciences seem to coalesce around the impacts on people or peoples' impacts on earth and its environments: climate change, global biodiversity, the impacts of fossil fuels, and so on. Other work in science in this area seems to focus on creating communities of (largely informatics) practitioners, or highlighting underrepresented voices in science (as much of the work in Wiki domains does). Finally, much scientific logic for this particular genre of linked data points to its ability to represent multiple forms of expressions at the same time – units and naming standards, data in multiple languages, or other divergent institutional or international conventions. Those efforts point to the well-known phenomenon in social studies of science that there is indeed a cultural aspect to how science is done, and how scientific knowledge is encoded into data, which in turn makes collections data difficult to reconcile across contexts (Delbourgo & Müller-Wille, 2012; Mee et al., 2022; Hjørland, 2022; Latour & Woolgar, 1986). The discovery of science's "dark data" across the world requires translation across human-produced knowledge infrastructures.

In the social sciences, the value propositions may be more obvious, but are perhaps still worth stating. For primary source repositories of import to social science, these collections and their embedded data are not just research evidence–they are also heritage, culture, and human knowledge. Thus, for many of the projects in these fields, linked data aspires to build connections across repositories and to generate impacts, not just for data and research fields, but for people and communities (facilitated, of course, by machines). Some projects, of course, go further—following the current CARE and Indigenous sovereignty principles, they seek to leverage linked data to return data to communities, or facilitate data discovery for its reclamation by communities. The projects we analyzed therefore largely seek to make collections data not only more discoverable, but more culturally relevant. Linked data can facilitate that. It allows multiple truths and forms of expressions at the same time–an important departure for cataloguing and representational systems that have long delimited (or worse, objectified) human knowledge into fixed categories (Moriarty & Turner, 2023; Mohan & Rodgers, 2021). On the other hand, Da Silva has noted that a major shift may be necessary for social scientists to move from holistic approaches to interpreting primary data to the "atomization" required of linked data, as well as the shift from manual to automated data collection, and the potential blurring of data and metadata (Da Silva 2018, 101-102). Yet, the social sciences are also more proximal to the humanities and to heritage work, both of which have built robust linked data infrastructures, as well as vocabularies, taxonomies, thesauri and other approaches to data models that may lay the groundwork for appropriate linked data ontologies.

While the sciences seem to have a lot of interest, the implementation of linked data seems to progress in fits and starts. It may be that the lift to embrace linked data in the sciences is too great relative to the anticipated benefits for various disciplines. Some fields, including earth sciences and biodiversity, have developed alternative approaches to data normalization and aggregation (some of which predate the emergence of linked data). While some approaches lack some benefits of a semantic approach, they have already gained a critical mass of users and implementers, making the shift to fundamentally different data models difficult. Linked data, like any infrastructure, relies on critical mass and the social cultures of knowledge domains. In this initial review, we chose to cut cases that merely built linked data add-ons or chose similar but alternative approaches. Ironically, many of those projects have stood

the test of time. Many of the projects we reviewed were online aggregators, databases, or portals created in the naughts and 2010s (all the rage!). Many have demonstrated real longevity. Rather than build new linked data infrastructures, many of these projects have built tools or resource pages to allow individual users or organizations to transform their data into linked-data-amendable formats. In scientific fields, where technological self-sufficiency is more common, that may be all that seems worth doing in the present moment. That approach may also be supported by the fact that many scientific databases and museum collections data are more inherently structured than, say, archival or special collections data, which (particularly in the finding aid tradition) is largely unstructured, and therefore needs far more concerted efforts to transform it for linked data platforms.

It is difficult to judge whether the sciences are rightly reluctant to jump on the linked data bandwagon. A major limitation of the present work is that it exposes only project ideals, not what linked data really are and are not good for in these various contexts. We know little about what of these aspirations has really been accomplished, nor whether linked data was in fact the best way to accomplish them. And of course, we know even less about the sustainability of any of this current work. We risk, like all other major innovations that have come before, betting on the wrong horse. We need to also ask ourselves: Is it the right answer for specific disciplines? What goals can be accomplished of this aspirational list without linked data? With what? And what is truly only possible via linked data? In the next phase, we aim to look more closely at disciplinary differences, the shape of evidence and knowledge in each domain, and their linked data implementations.

CONCLUSION

Critical studies of the processes of archival datafication – the derivation of digital data from primary sources – have been siloed by disciplines, and focus predominantly on humanistic applications. This is a missed opportunity for understanding (a) the evolution of scientific and social scientific research practices, which have interpretive components in relation to primary sources; and (b) deepening understanding of the ethics of datafication and especially linked datafication in a broader range of contexts.

To date, very little is understood about how primary sources are used in sciences and social sciences, or why and how these sources are being transformed into linked data representations. In this paper, we have therefore shared the findings of an empirical attempt to understand how and why and when professionals and researchers in the sciences and social sciences leverage linked data for analog primary source collections. At this stage, we have already begun to identify some disciplinary divergences among the use of primary sources, their conception as data, and their integration with linked data. Our work to select and code cases illuminated the blurry nature of many of these core categories–primary sources, disciplinary boundaries, and what constitutes linked data. Further, we have more work to do to understand the realities on the ground, and the affordances of linked data relative to other possibilities.

Our future work will apply some of the linked data methods and workflows we documented here with test collections in anthropology (a field that cuts across the social sciences and sciences) to understand what is gained and lost in various data representations and linked data approaches. As a touchstone, we aim to leverage anthropology's inherent interdisciplinarity to both look and speak broadly through the work. Additionally, we plan to review particularly relevant linked data projects in fields external to the sciences and social sciences to generate a more comprehensive picture of linked data as documented in cultural heritage literature.

Perhaps these initial findings also suggest that we should test other alternatives as well. Beyond the extant possibilities, it is important to note that alternative approaches to data management (data mesh, AI, and active metadata approaches, for example) are emergent areas for innovation in this domain that need exploration.

Throughout this project, we aim to maintain a grounded understanding that will allow us to capture the evolving shape of knowledge representation and research practices, and their potential impacts on the future of scientific, scholarly, and community records. This cross-disciplinary approach will help to break down silos between fields, and with benefits LAM and data curation professional practice in domains new to these discussions.

GENERATIVE AI USE

Generative AI was not used in this research or authorship of this publication.

AUTHOR ATTRIBUTION

First Author: supervision, conceptualization, funding acquisition, project administration, methodology, formal analysis, writing – inc. original draft, review, and editing; Second Author: supervision, conceptualization, funding acquisition, project administration, methodology, formal analysis, writing – inc. original draft, review, and editing; Third Author: funding acquisition, investigation, formal analysis, project administration, writing – inc. original draft, review, and editing; Fourth Author: investigation, formal analysis, writing – inc. original draft, review, and editing.

ACKNOWLEDGMENTS

We thank the National Science Foundation Cultural Anthropology program for their generous support of this research (NSF CA-SR award #2314762). We are enormously grateful to our Council for the Preservation of Anthropological Records (CoPAR) Working Group, who have workshopped this project and its methods with us at key stages, and to CoPAR's Advisory Board, whose members provided important feedback as we drafted our grant proposal. At the University of Maryland, we acknowledge the work of Maura Matvey, Polly O'Rourke, and Susan Winter for their grant-writing and development support. We also thank the anonymous reviewers whose insightful comments shaped our final paper.

REFERENCES

Atlas of Living Australia (n.d.). Open access to Australia's biodiversity data. Retrieved April 9, 2024, from https://www.ala.org.au/

Archaeology Data Service. (n.d.). The digital repository for archaeology and heritage, supporting access, innovation, and research. Retrieved April 9, 2024, from https://archaeologydataservice.ac.uk/

Arkisto. (n.d.). Case Study: Modern PARADISEC. Retrieved April 9, 2024, from https://arkisto-platform.github.io/case-studies/paradisec/

Berners-Lee, T. (2006, July 27). *Linked Data—Design Issues*. https://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*. https://www-sop.inria.fr/acacia/cours/essi2006/Scientific%20American_%20Feature%20Article_%20The%20Semantic%20Web_%20May%202001.pdf

Binding, C., Evans, T., Gilham, J., Tudhope, D. and Wright, H. (2022). Linked Data for the Historic Environment. *Internet Archaeology, 59*. https://doi.org/10.11141/ia.59.7

Bio2RDF v2.7a. (n.d.). Linked Data for the Life Sciences. Retrieved April 9, 2024, from https://bio2rdf.org/

Bionomia. (n.d.). Rationale. Retrieved April 9, 2024, from https://bionomia.net/about

Biodiversity Collections Network. (n.d.). Building a community to advance and sustain digitized biocollections. Retrieved April 9, 2024, from https://bcon.aibs.org/

Biodiversity Heritage Library. (n.d.). Unifying Biodiversity Knowledge to Support Life on a Sustainable Planet. Retrieved April 9, 2024, from https://bhl.pubpub.org/

Brunet, M., & Jones, P. (2011). Data rescue initiatives: Bringing historical climate data into the 21st century. *Climate Research*, *47*(1/2), 29–40.

Buchanan, S. A. (2019). The assemblage of repository and museum work in archaeological curation. *Information Research, 24*(2).

Callahan, A., Cruz-Toledo, J., & Dumontier, M. (2013). Ontology-Based Querying with Bio2RDF's Linked Open Data. *Journal of Biomedical Semantics*, *4*(S1), S1. https://doi.org/10.1186/2041-1480-4-S1-S1

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE Principles for Indigenous Data Governance. *Data Science Journal*, *19*, 43. https://doi.org/10.5334/dsj-2020-043

Cliggett, L. (2013). Qualitative Data Archiving in the Digital Age: Strategies for Data Preservation and Sharing. *The Qualitative Report*, *18*(24), 1–11.

Cornut, M., Raemy, J. A., & Spiess, F. (2023). Annotations as Knowledge Practices in Image Archives: Application of Linked Open Usable Data and Machine Learning. *Journal on Computing and Cultural Heritage*, *16*(4), 80:1-80:19. https://doi.org/10.1145/3625301

Dallas, C. (2016). Digital curation beyond the "wild frontier": a pragmatic approach. *Archival Science 16*, 421–457. https://doi.org/10.1007/s10502-015-9252-6

Daniels, M. (2014). *Data Reuse in Museum Contexts: Experiences of Archaeologists and Botanists* [Doctoral Dissertation, University of Michigan].

DarwinCore. (n.d.). DarwinCore. Retrieved April 9, 2024, from https://dwc.tdwg.org/

Da Sylva, L. (2018). Towards linked data: Some consequences for researchers in the social sciences and humanities. *Proceedings of the Association for Information Science and Technology, 55*(1), 94-103.

Davis, E., & Heravi, B. (2021). Linked Data and Cultural Heritage: A Systematic Review of Participation, Collaboration, and Motivation. *Journal of Computing in Cultural Heritage, 14*(2), 1-18. https://doi.org/10.1145/3429458

Delbourgo, J., & Müller-Wille, S. (2012). Introduction (Focus: Listmania). *Isis, 103*(4), 710–715. https://doi.org/10.1086/669045

DigIn. (n.d.). Project Description. Retrieved April 9, 2024, from https://www.digin-tcn.org/about/project-description

Distributed System of Scientific Collections. (n.d.). Distributed System of Scientific Collections. Retrieved April 9, 2024, from https://www.dissco.eu/

EBI. (n.d.) EBI Search RESTful Web Services. Retrieved April 9, 2024, from https://www.ebi.ac.uk/ebisearch/documentation/rest-api

Europeana. (n.d.). About. Retrieved April 9, 2024, from https://www.europeana.eu/en/about-us

Flensburg, S., & Lomborg, S. (2023). Datafication research: Mapping the field for a future agenda. *New Media & Society, 25*(6), 1451-1469. doi:10.1177/14614448211046616

Furner, J. (2004). Conceptual Analysis: A Method for Understanding Information as Evidence, and Evidence as Information. *Arch Sci 4*, 233–265. https://doi.org/10.1007/s10502-005-2594-8

Gaitanou, P., Andreou, I., Sicilia, M.-A., & Garoufallou, E. (2024). Linked data for libraries: Creating a global knowledge space, a systematic literature review. *Journal of Information Science*, *50*(1), 204–244. https://doi.org/10.1177/01655515221084645

GBIF. (n.d.). What is GBIF? Retrieved April 9, 2024, from https://www.gbif.org/what-is-gbif

GeoCASe. (n.d.). *GeoCASe 2.0: The Earth Science Collections Portal*. Retrieved April 9, 2024, from https://geocase.eu/

Geser, G. (2016). *WP15 Study: Towards a Web of Archaeological Linked Open Data*. Salzburg Research.

Granikov, V., Hong, Q. N., & Pluye, P. (2022). Mixing Qualitative and Quantitative Evidence in a Systematic Review: Methodological Guidance With a Worked Example of Collaborative Information Monitoring. In P. Ngulube (Ed.), *Advances in Knowledge Acquisition, Transfer, and Management* (pp. 125–146). IGI Global. https://doi.org/10.4018/978-1-7998-8844-4.ch007

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*, *26*(2), 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Hajra, A., & Tochtermann, K. (2017). Linking science: Approaches for linking scientific publications across different LOD repositories. *International Journal of Metadata, Semantics and Ontologies*, *12*(2–3), 124–141. https://doi.org/10.1504/IJMSO.2017.090778

Hasnain, A., Mehmood, Q., Sana E Zainab, S., Saleem, M., Warren, C., Zehra, D., Decker, S., & Rebholz-Schuhmann, D. (2017). BioFed: Federated query processing over life sciences linked open data. *Journal of Biomedical Semantics*, *8*(1), 13. https://doi.org/10.1186/s13326-017-0118-0

Hawkins, A. (2022). Archives, linked data and the digital humanities: Increasing access to digitised and born-digital archives via the semantic web. *Archival Science, 22*(3), 319–344. https://doi.org/10.1007/s10502-021-09381-0

Henare, A., Holbraad, M., & Wastell, S. (2007). *Thinking Through Things: theorising artefacts ethnographically*. New York: Routledge.

Hjørland, B. (2022). Science, Part II: The Study of Science. *Knowledge Organization, 49* (4), 273–300. https://doi.org/10.5771/0943-7444-2022-4-273

Hyvönen, E. (2012). *Publishing and using cultural heritage linked data on the semantic web* (Vol. 3): Morgan & Claypool Publishers.

IDEA (n.d.). IDEA: International (Digital) Dura-Europos Archive. Retrieved April 9, 2024, from https://duraeuroposarchive.org/

IDigBio (n.d.). Home. Retrieved April 9, 2024, from https://www.idigbio.org/

Jaillant, L. (2022). How can we make born-digital and digitised archives more accessible? Identifying obstacles and solutions. *Archival Science* 22, 417–436. https://doi.org/10.1007/s10502-022-09390-7

Jentzsch, A., Zhao, J., Hassanzadeh, O., Cheung, K-H., Samwald, M., & Andersson, B. (2009). Linking open drug data. In A. Paschke, H. Weigand, W. Behrendt, K. Tochtermann, & T. Pellegrini (Eds.), 5th International Conference on Semantic Systems, Graz, Austria, September 2-4, 2009. Proceedings. Verlag der Technischen Universität Graz. http://i-semantics.tugraz.at/2009/triplification/05_TriplificationChallengeLODD.pdf

Johnson, K. R., Owens, I. F. P., & the Global Collection Group. (2023). A global approach for natural history museum collections. *Science*, *379*(6638), 1192–1194. https://doi.org/10.1126/science.adf6434

Johnston, L. R., Carlson, J., Hudson-Vitale, C., Imker, H., Kozlowski, W., Olendorf, R., Stewart, C., Blake, M., Herndon, J., McGeary, T. M., & Hull, E. (2018). Data Curation Network: A Cross-Institutional Staffing Model for Curating Research Data. *International Journal of Digital Curation*, *13*(1), Article 1. https://doi.org/10.2218/ijdc.v13i1.616

Jones, L. V., & Coggeshall, P. (Eds.). (1982). *An Assessment of Research-Doctorate Programs in the United States: Social and Behavioral Sciences*. National Academy Press. https://doi.org/10.17226/9781

Kalfatovic, M.R., Crowley, B., Dearborn, J., Funkhouser, C., Iggulden, D., Trei, K., Herrmann, E., and Merriman, K. (2023). Safeguarding Access to 500 Years of Biodiversity Data: Sustainability planning for the Biodiversity Heritage Library. Biodiversity Information Science and Standards 7, e112430. https://doi.org/10.3897/biss.7.112430

Kansa, S. W. (2015). Using Linked Open Data to Improve Data Reuse in Zooarchaeology. *Ethnobiology Letters*, *6*(2), 224–231. http://www.jstor.org/stable/26423627

Kelly, Julia A., Farrell, Shannon L., Hendrickson, Lois G., Luby, J., Mastel, K. L. (2022). A Critical Literature Review of Historic Scientific Analog Data: Uses, Successes, and Challenges. *Data Science Journal, 19*(14), 1-11. https://doi.org/10.5334/dsj-2022-014

Kirakosian, K., & Bauer-Clapp, H. (2017). A Walk in the Woods: Adapting Archaeological Training to Archival Discovery. Advances in Archaeological Practice, 5(3), 297–304. https://doi.org/10.1017/aap.2017.17

Kouper, I. (2016). Professional participation in digital curation. *Library & Information Science Research, 38*(3), 212–223. https://doi.org/10.1016/j.lisr.2016.08.009

Kriesberg, A., Frank, R. D., Faniel, I. M., & Yakel, E. (2013). The role of data reuse in the apprenticeship process. *Proceedings of the American Society for Information Science and Technology*, *50*(1), 1–10. https://doi.org/10.1002/meet.14505001051

Latour, B., & Woolgar, S. (1986). *Laboratory life: The construction of scientific facts*. Princeton University Press.

Marsh, D. E., St. Andre, S., Wagner, T., & Bell, J. A. (2023). Attitudes and uses of archival materials among science-based anthropologists. *Archival Science, 23*(3), 355-379. doi:10.1007/s10502-023-09411-z

Mayernik, M. (2016). Research data and metadata curation as institutional issues. *Journal of the Association for Information Science and Technology*, 67(4): 973–993. https://doi.org/10.1002/asi.23425

Mee, J., Sangster, M., & Porter, D. (Eds.). (2022). Catalogues as Instituting Genres of the Nineteenth Century Museum: The Two Hunterians. In *Institutions of Literature, 1700–1900* (1st ed., pp. 157–177). Cambridge University Press. https://doi.org/10.1017/9781108909501

Mohan, U., & Rodgers, S. (2021). Classification Schemes Gone Awry: Implications for Museum Research and Exhibition Display Practices. *Museum Anthropology, 44*(1–2), 4–10. https://doi.org/10.1111/muan.12238

Moriarty, P., & Turner, H. (2023, January 1). Junk and Priceless China: A Chronology of Cataloging at the Museum of Anthropology. CAIS2023. https://cais2023.ca/talk/18.moriarty/

National Center for Education Statistics. (n.d.). The Classification of Instructional Programs. Retrieved April 9, 2024, from https://nces.ed.gov/ipeds/cipcode/cipdetail.aspx?y=55&cipid=87817

Nelson, A., 2022. Memorandum for the heads of executive departments and agencies: Ensuring free, immediate, and equitable access to federally funded research. Retrieved April 9, 2024, from https://www.whitehouse.gov/ostp/news-updates/2022/06/21/readout-of-dr-alondra-nelsons-participation-in-the-g7-science-ministerial-progress-toward-a-more-open-and-equitable-world/

Nundloll, V., Smail, R., Stevens, C., & Blair, G. (2022). Automating the extraction of information from a historical text and building a linked data model for the domain of ecology and conservation science. Heliyon, 8(10), e10710. https://doi.org/10.1016/j.heliyon.2022.e10710

Nurmikko-Fuller, T. (2024). Linked Data for Digital Humanities (1–1 online resource (xxiv, 129 pages).). Routledge. https://www.taylorfrancis.com/books/9781003197898

Open Context. (n.d.). Web Services (APIs). Retrieved April 9, 2024, from https://opencontext.org/about/services

Open GLAM. (n.d.). Open GLAM: A global network on sharing cultural heritage. Retrieved April 9, 2024, from https://openglam.org/

Padilla, T., Scates Kettler, H., & Shorish, Y. (2023). Collections as Data: Part to Whole Final Report. Zenodo. https://doi.org/10.5281/zenodo.10161976

Paulus Jr., M. J. (2012). What is primary: Teaching archival epistemology and the sources continuum. In E. Mitchell, P. Seiden, & S. Taraba (Eds.), *Past or Portal?: Enhancing Undergraduate Learning Through Special Collections and Archives* (pp. 76–82). Association of College and Research Libraries.

PubChem. (n.d.). PubChem. Retrieved April 9, 2024, from https://pubchem.ncbi.nlm.nih.gov/

PubMed. (n.d.). PubMed. Retrieved April 9, 2024, from https://pubmed.ncbi.nlm.nih.gov/

Saleem, M., Padmanabhuni, S. S., Ngonga Ngomo, A.-C., Iqbal, A., Almeida, J. S., Decker, S., & Deus, H. F. (2014). TopFed: TCGA Tailored Federated Query Processing and Linking to LOD. *Journal of Biomedical* Semantics, 5(1), 47. https://doi.org/10.1186/2041-1480-5-47

Samwald, M., Jentzsch, A., Bouton, C., Kallesøe, C. S., Willighagen, E., Hajagos, J., Marshall, M. S., Prud'hommeaux, E., Hassanzadeh, O., Pichler, E., & Stephens, S. (2011). Linked open drug data for pharmaceutical research and development. *Journal of Cheminformatics, 3*(1), 19. https://doi.org/10.1186/1758-2946-3-19

Schmidt SC, Thiery F., and Trognitz M. (2022). Practices of Linked Open Data in Archaeology and Their Realisation in Wikidata. *Digital 2*(3), 333-364. https://doi.org/10.3390/digital2030019

Sorensen, Amanda. (2023). Reflections: Interoperating and Aggregating Natural History Collections. *Axiell Blog,* December 11, 2023. https://www.axiell.com/blog-post/reflections-interoperating-and-aggregating-museum-collections/

Sorensen, A. H., Escobar-Vredevoogd, C., Wagner, T. L., & Fenlon, K. (2023). Recovering and Reusing Historical Data for Science: Retrospective Curation Practices Across Disciplines. In I. Sserwanga, A. Goulding, H. Moulaison-Sandy, J. T. Du, A. L. Soares, V. Hessami, & R. D. Frank (Eds.), *Information for a Better World: Normality, Virtuality, Physicality, Inclusivity* (pp. 14–28). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-28035-1_2

Thomer, A. K. (2022). Integrative data reuse at scientifically significant sites: Case studies at Yellowstone National Park and the La Brea Tar Pits. *Journal of the Association for Information Science and Technology*, 73 (8), 1155–1170. https://doi.org/10.1002/asi.24620

Tudhope, D., Jeffrey, S., Binding, C. (n.d.) Semantic Technologies Enhancing Links and Linked data for Archaeological Resources (STELLAR). Retrieved April 9, 2024, from https://gtr.ukri.org/projects?ref=AH%2FH037357%2F1#/tabOverview

van Hoolan, S., & Verborgh, R. (2014.) *Linked Data for Libraries, Archives, and Museums.* London: Facet Publishing.

W3C. (n.d.). W3C Glossary and Dictionary. Retrieved April 9, 2024, from https://www.w3.org/2003/glossary/

GBIF. (n.d.). What is GBIF? Retrieved April 9, 2024, from https://www.gbif.org/what-is-gbif

Walter, M., & Suina, M. (2019). Indigenous data, Indigenous methodologies and Indigenous data sovereignty. *International Journal of Social Research Methodology, 22*(3), 233-243.

Wiedeman, G. (2023). Designing Digital Discovery and Access Systems for Archival Description. *Code4Lib Journal*(55). Retrieved from https://journal.code4lib.org/articles/16963

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., …Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Willighagen, E. L., Waagmeester, A., Spjuth, O., Ansell, P., Williams, A. J., Tkachenko, V., Hastings, J., Chen, B., & Wild, D. J. (2013). The ChEMBL database as linked open data. *Journal of Cheminformatics*, *5*(1), 23. https://doi.org/10.1186/1758-2946-5-23

Wippich, Carol. (2012). Preserving science for the ages—USGS data rescue. *U.S. Geological Survey Fact Sheet 2012–3078*, 4. https://pubs.usgs.gov/fs/2012/3078.

Yang, Y. (2024). Datafication of audiovisual archives: from practice mapping to a thinking model. *Journal of Documentation, ahead-of-print*(ahead-of-print). doi:10.1108/JD-04-2022-0093

Zapilko, B., Schaible, J., Wandhöfer, T., and Mutschke, P. (2016). Applying Linked Data Technologies in the Social Sciences. *Künstl Intell 30*, 159–162. https://doi.org/10.1007/s13218-015-0416-6

Zhang, Y., & Wildemuth, B. M. (2016). Qualitative analysis of content. In *Applications of social research methods to questions in information and library science* (pp. 318–329). ABC-CLIO, LLC. https://www.ischool.utexas.edu/~yanz/Content_analysis.pdf