Feast and Famine: The Problem of Sources for Linked Data Creation

Rebecca Kahn[1] ![ORCID]

Rainer Simon[2] ![ORCID]

[1] Alexander von Humboldt Institute for Internet and Society
Berlin, Germany
[2] AIT Austrian Institute of Technology
Vienna, Austria

Abstract

In this article, we reflect on some of the challenges that are encountered when applying Linked Data within the context of museum collections. On the one hand, we discuss the more ostensible challenges of data integration, vocabulary mapping, licensing, and attribution. On the other, we seek to stimulate a wider discussion of the complexities of publishing museum data (which often contains multiple layers of provenance and copyright information, as well as rich contextual metadata) using a combination of the triple-based structure of RDF and standard ontologies. We will show that while ideal best-practice applications do exist, they are not always used in the case of museum data 'in the wild.' This disjuncture has significant implications for the responsible publication of museum data, especially in the light of recent efforts within the museum world to decolonize collections by reconsidering the provenance and historical contexts of objects and data.

1   Introduction

The creation of annotations is one of the basic functions of scholarly prac-
tice across disciplines (Blanke and Hedges, 2013). In the digital space, such
annotations can emulate analog practices, assuming the role of note-taking,
summarizing, or highlighting, but they can also go much further and ex-
tend existing practice by becoming a mechanism for collaboration and the
sharing of information. Semantic annotation (i.e. the marking up of text
and images with references to controlled vocabularies) enables the linking of
annotations to each other and to secondary sources, stored in repositories
and collections across the web. In this paper, we reflect on the challenges
and opportunities of linking online collections by means of digital annota-
tion. We take a two-part approach based on our experience of working with
Linked Open Data (LOD) annotations of historical sources. We start with
a practical perspective, and consider the technical and methodological dif-
ficulties of enabling these linkages, in particular when creating direct links
between objects and documents. We then explore the theoretical implica-
tions of this method, and ask whether the ability to link objects should be
tempered by questions of data provenance and data models. These questions
will be framed by reference to museum collections, which do not always lend
themselves easily to the interconnectedness and openness required to make
LOD useful.

2   Background

Pelagios (Isaksen et al., 2014) is an international digital humanities network,
which facilitates the linking of online resources documenting the past via
the place names that occur in them. In 2014, Pelagios created Recogito,[1]
an online environment for the geographic annotation of historical sources
(Simon et al., 2019). Fundamental to Recogito is a graph model that al-
lows scholars to annotate sources with Linked Data identifiers provided by
gazetteers, thus building connections between documents that refer to the
same places. This tool has been enthusiastically taken up by scholars, who
have been able to transfer their analog practice of annotation into the digital
sphere, and extend it by leveraging the Linked Data connections to navig-
ate between sources that share relations to place. One of the aspects which
users appreciate most is the ease of use: an essential goal of the developers
was to create a tool that was sufficiently generic to cater to different research
needs, while not being engineered beyond the requirements of the average
researcher, who may not have much interest in how Recogito works, as long

---

[1] https://recogito.pelagios.org

as it does.

As the user base has grown, we have increasingly realized how the ability to create useful connections crucially depends on the availability of suitable linked data vocabularies and name authorities. With regard to geographic annotation, for example, Recogito can only be truly useful to a specific research community if it integrates a critical mass of the right gazetteers – linked data authorities that cover that community's domain appropriately in terms of geographic extent, granularity, temporal and cultural focus, etc. While such scholarly name authorities have been successfully established in some communities (e.g. in the Classics, with the Pleiades gazetteer[2]), other communities (e.g. scholars of Early Modern history) lack established scholarly gazetteers. There is a variety of reasons for this discrepancy, some of which are rooted in the period in question; for example, as McDonough and van der Camp point out, in early modern France, place names were in political and geographical flux in the years before and after the French revolution, making it difficult to match pre- and post-Revolution counterparts (McDonough and van der Camp, 2017). In other cases, a lack of technical resources, such as Natural Language Processing tools for particular languages, or textual geoparsing tools which are limited to particular historical periods or regions (Murrieta-Flores and Gregory, 2015) have restricted the development of authoritative gazetteers. Without these, scholars wanting to use Recogito are left with two options: either they work with a set of generic place names and LOD identifiers provided by online sources such as GeoNames[3] or Wikidata,[4] or they bootstrap their own customized gazetteers, thus somewhat defeating the purpose of using Linked Data in the first place. Finding ways to link people, events, or sources is equally, if not more, challenging.

In addition to the problem of Linked Data authorities, there remains a question about the model of representing the data that we aim to link through them. The Resource Description Framework (RDF) is the fundamental "markup language" for encoding Linked Data. But is its minimal triple structure helpful as a foundation and mental model when conceptualizing humanities data? Challenges in this regard relate, among others, to the issue of resource identity, and to the distinction between information and non-information resources. Born digital material may, or may not, have a physical counterpart. If they do, they may carry hidden assumptions and biases in the way that they represent or describe physical items, or agglomerate information from multiple information sources without making it ex-

---

[2] https://pleiades.stoa.org/

[3] https://www.geonames.org/

[4] https://www.wikidata.org/

plicit. While the "shape" of Linked Data is ideal for pulling together these multiple sources and perspectives, the subject-predicate-object construction of the RDF triple does not make it easy to include the essential contextualizing data that may accompany these information objects; at least not in a straightforward manner. Bechhofer et al. (2013) describe in detail how in medical research the publication of triples without contextualizing data may be counterproductive to the development of scientific research methodologies, including the provision of provenance data, which aids the interpretation and trust of results, and methods to support reproducibility. They argue that publishing results as LOD poses the risk of confounding the flow of rights to the researcher. Although we are dealing with different material in this paper, it is possible to ask similar questions when considering the use of RDF as a tool for representing and publishing some types of cultural heritage sources. Over the following sections, we will describe some of the contexts of humanities scholarship in which it is appropriate to ask these questions, and explore the implications of this for digital humanities work.

3   The Feast

The digitization of cultural heritage data is creating previously unimagined possibilities for humanities researchers. Since the Semantic Web was first presented as a theoretical concept in the late 1990s (Berners-Lee, 1999), cultural heritage institutions have been quick to realize the transformative potential offered by Linked Data and semantic modeling, recognizing it as a way to create connections between previously siloed collections of data. Museum, archive, and library materials are now available online, and the growth of large-scale digital infrastructure projects such as Europeana[5] and the European Holocaust Research Infrastructure[6] allow scholars to overcome many of the logistical challenges of using primary sources, such as the fragmentation and geographic dispersal of collections. Collaborative projects such as Nomisma.org have linked hundreds of thousands of archival records and knowledge objects into networks and infrastructures, thereby enabling comprehensive views of previously disparate collections. The success of these initiatives, however, should not mask the difficulty involved in transforming collections of data that were previously stored in relational databases into knowledge graph models, or the compromises between interoperability and specificity that need to be made with regards to the data itself. Furthermore, it remains unclear whether the minimal triple structure of RDF is a good starting point when conceptualizing some types of humanities data,

---

[5] https://classic.europeana.eu/portal/en
[6] https://www.ehri-project.eu/

specifically sources that are stored in cultural heritage collections such as galleries, libraries, archives, and museums (GLAMs). While the triple model facilitates the creation of complex networks, and shared authority files enable the opening up of silos of data, the question is whether such a model prioritizes connections over complexity, and what the middle ground between these two data ideals might be.

3.1 Difficulties in Mapping Data

Mapping museum data to RDF is a considerable task. Many large museums have long histories of acquisition, which has often resulted in idiosyncratic record-keeping and documentation practices (Blagoev et al., 2018). These practices have produced legacy data, which is often unique to individual institutions and in some cases to individual catalogers within those institutions, with different approaches to documentation being employed across different curatorial departments and object types. The British Museum, for example, holds artworks by Rembrandt in the same collections management system as multiple terracotta lamps created by unnamed artisans in antiquity. These are different objects, which require different approaches to describe and record their details and the knowledge that has emerged from their study. In terms of content, museum data may be messy as a result of many years of labor, mistakes, and revisions by many different hands. Many collections feature text fields which contain a great deal of unstructured metadata in the form of free text strings, which provide additional data about an object, its provenance, or associated individuals and groups. Finally, when mapping museum data onto an ontology, it is not uncommon to be faced with data that arrives in a range of different formats, from CSV to XML and JSON (Knoblock et al., 2017).

As discussed above, the move from a flat, relational data model to a graph with the ability to supplement and extend information objects by expressing complex relationships within the data offers great promise. However, getting the data to fit into such a format often involves extensive processes of standardizing and cleaning. Standardization into ontologies or controlled vocabularies is vital for the transformation of the data in question, but, as Kelly Davis argues, it is a valid process only if the result remains usable within the Linked Data model (Davis, 2019). This argument is echoed by the experience of the American Art Collaborative project,[7] which used Karma,[8] an intermediary tool, to map the data from 14 art museums to Linked Data, using the de facto standard museum ontology, the CIDOC Conceptual Reference

---

[7]https://americanart.si.edu/about/american-art-collaborative
[8]https://usc-isi-i2.github.io/karma/

90

Model (CIDOC CRM)[9]. In this case, there was no consensus between the CIDOC experts on the project as to how certain data should be mapped, which resulted in a suspension of the mapping until agreement could be reached (Knoblock et al., 2017). This complexity, and a paucity of expertise among museum staff to resolve it, is often a contributing factor to the overall lack of museum data in the Linked Data ecosystem (Geddes, 2019).

When developing Recogito, we solved the problem of making data that differ vastly in terms of size, content type, and theme uniformly accessible under a single user interface by taking a pragmatic approach. Realizing that it was unlikely that we would be able to convince our partners to all agree on one method of representing their data, Pelagios provides a set of light-weight conventions for how to express links between the data and the things described in it, which in most cases were geographic descriptions. We refer to this approach as "connectivity through common references" (Simon et al., 2019). However, this solution can only be operationalized if there is already a critical mass of material available as LOD, which is not yet the case when it comes to ethnographic museum data.

4   Changing Data

Unlike library records, which provide the metadata that facilitates access to an information object (Gilliland, 2016), museum records typically involve a great deal of contextualizing and descriptive information at the individual item level, including descriptive data, rights data, as well as technical and preservation data. Typically, a museum object record will also be the place where any new research about the object is recorded, including its exhibition history. In this way, it is fair to say that a museum record is never truly complete, since the information contained within it is liable to be constantly changed and updated. In the context of ethnographic museum collections, this presents a particular set of challenges. One significant tension that has to be managed is between the volume of data and the individual, item-level object biographies. Scale is an essential aspect of Linked Data,[10] since the model is designed to work best when there are multiple connections across many different data repositories. This system is dependent on shared terminology, usually mapped to an authority file, which allows terms that are semantically similar to be recognized, and the linkages to be made. However, the historical nature of these materials and their documentation means that even in the analogue versions, there is often very little standardization across fields. This is particularly evident in the notes fields of many collec-

---

[9]http://www.cidoc-crm.org/
[10]https://www.w3.org/standards/semanticweb/data

tions, which contain information on the object itself, details about provenance, or extensive notes provided by the curators, who are often subject experts (Griffiths, 2010). This data presents a conundrum: either it can be excluded from what is visible in the aggregator (as is the case with Recogito), although it may still be found by following the links to the original source record in the data provider's repository; or it has to be manually remodeled in order to be included, as was the case with the American Art Collaborative. Both options have distinct benefits and drawbacks.

5 Managing the Mass

In recent years, there has been an increasing discussion among museum scholars and professionals about the need to decolonize their collections. In many cases, these discussions focus on individual objects, their display and documentation, although they also have an effect on the data infrastructures. Inclusive terminologies, which reflect the practices of the communities in which the objects originated, are being added to many records and cultural sensitivity warnings are becoming increasingly ubiquitous when accessing collections databases online.[11] Many of these efforts are codified in the 2013 International Council of Museums Code of Ethics, which provides recommendations for the acquisition, storage, research, and display of culturally sensitive material. But in the context of converting museum records to Linked Data, in which an institution provides access to many thousands of records via a SPARQL endpoint or a data dump, the practical realities of managing changes like this can be a challenge, particularly if the revisions are being made retrospectively, after the LOD workflow has been implemented.

One illustrative example is a selection of records of human remains originating from what was then the Belgian Congo (today the Democratic Republic of Congo), which are now part of the collection of the Museum of Ethnography in Stockholm. The museum's collections are aggregated by the national LOD aggregator Swedish Open Cultural Heritage (SOCH),[12] which currently contributes over 2,400,000 records from around 45 institutions to Europeana.[13] A search by contributing institutions reveals that the Ethnographic Museum has contributed 238,138 objects to Europeana via SOCH. Within this subset, a search using the terms 'human remains' (*mänskliga kvarlevor* in Swedish) renders 1,010 results. Of these, the majority of

---

[11]See, for example, the cultural warnings for anthropology museums such as the Pitt Rivers Museum in Oxford
(`https://prm.web.ox.ac.uk/terms-use-pitt-rivers-museum-database-object-collections`) or the South Australia Museum (`https://www.samuseum.sa.gov.au/cultural-sensitivity-warning`).

[12]`https://www.raa.se/in-english/digital-services/about-soch/`

[13]`https://pro.europeana.eu/organisation/swedish-open-cultural-heritage`

the records display a grayed-out image panel, which states 'Human Remains, image has been blocked.' However, six items do have visible accompanying images, which show human skulls in various states of completeness, some adorned with shells and feathers. All six of the images and records are freely available for reuse, with Creative Commons CC BY licences, which require only attribution of the original source, or CC 0 (public domain) licenses. No contextualizing information on how the materials came to be in the collection, where, when, and under what circumstances they were obtained, is to be found in the accompanying metadata. At no point in the record is there a statement from either Europeana or the museum which acknowledges that these are human remains, addresses how they are displayed or documented, or explains why they are visible, when other items in the subset are not. This is despite the fact that the ICOM Code of Ethics, subsection 4.3, states that: "Human remains and materials of sacred significance must be displayed in a manner consistent with professional standards and, where known, taking into account the interests and beliefs of members of the community, ethnic or religious groups from whom the objects originated. They must be presented with great tact and respect for the feelings of human dignity held by all peoples."[14] This example represents a very small percentage of the overall number of objects from the museum's collection which have been added to Europeana, and is not meant to imply that the Ethnographic Museum, Swedish Open Cultural Heritage, or Europeana are knowingly sharing images of human remains on their websites or portals. Rather, what they illustrate is the complexity of managing large sets of heterogeneous data characterized by a significant variance in their descriptive terminologies, and the need for added precautionary measures when certain types of material are part of those datasets.

6   The Famine

The famine in the title of this paper not only refers to the lack of gazetteers as a source for creating semantic links between sources, but also to the inherent difficulties of managing large sets of data, which, ironically, can result in a dearth of contextualized, rich, and usable Linked Data that is at the disposal of researchers.[15] The cultural heritage sources which form the basis of much of humanities scholarship are diverse and complex. Digitized sources may include transcribed or OCRed texts, scanned images, high resolution photographs or stills, musical scores, and census or survey data. The digitization process may enrich these knowledge objects with supplementary metadata,

---

[14]https://icom.museum/en/resources/standards-guidelines/code-of-ethics/
[15]https://linked.art/loud/

drawn from a range of different sources. Meanwhile, born digital sources (such as the annotations in Recogito) may be completely new or only vaguely resemble the original, as they contain an agglomeration of multiple information sources. This re-aggregation of data into a triple means that by default, some data will not be automatically included in the digital source. Persistent provenance data, for example, which aids the interpretation and trust of results, and facilitates methods to support reproducibility, is not natively included in the model (Sikos and Philip, 2020). This omission has significant implications for linked cultural heritage data, where provenance information is essential for providing context to the object, as is particularly evident in the case of ethnographic collections, as well as collections containing objects that may be considered to be looted or stolen art.

6.1 Copyright and Permission Problems

One of the minimal conceptual drivers behind the Semantic Web is that the data being shared is openly licensed, thus enabling the linking and sharing of datasets in non-proprietary formats. As illustrated in the example provided above, even this most basic of assumptions is complicated in the context of museum data collections, which often contain a range of different copyright statements within one collection. In many museum collections, it is extremely difficult to identify the copyright status of individual collection items, and identifying rights holders in order to obtain clearance can be complex and time consuming, if not impossible (Wallace and Euler, 2020). Limited or incomplete information is often all that is available for works, or they may have multiple rights holders. These conditions make the management of copyright data within an LOD framework extremely complex, if, for example, not all of the data within a dataset can be openly licensed without risking a contravention of the copyrights pertaining to them. A 2018 study (Blijden, 2018) on the accuracy of the rights statements in collections included in Europeana found that 19% of the collections examined (a total of 10 collections) had inaccurate `edm:rights` values, while accuracy for a further 25% (15 collections) could not be determined. Bearing in mind that each collection contains many hundreds of thousands of objects, this means that a significant set of data has been released into the web with inaccurate or incomplete rights information. The researchers also found that, in general, the more heterogeneous the collections, the more likely the data was to be incomplete or inaccurate. This example highlights the complexities involved in managing important metadata for large collections of linked open cultural heritage data, and the risks posed by releasing material for which there is missing or incomplete copyright documentation.

6.2 Ethnographic Collections

The recent move towards critical self-reflection in museum practice has highlighted the need for museums to both engage with and communicate how their collections came into being. Many of the world's great encyclopedic museums were created as part of larger national and imperial expansion projects during the nineteenth and early twentieth centuries (Turner, 2015). The ways in which material was collected, cataloged, and described, but also the language and terminology included in the collections records, reflect these historical realities: ethnographic objects were seen as scientific specimens rather than individual works of art or human remains, and the biographies of those objects and their creators often went unrecorded (Beltrame, 2016). Today, many collections around the world are beginning to recognize and acknowledge these histories, and the role that plunder, coercion, and political violence may have played in their formation. Efforts are now being undertaken in institutions around the world to revise these collections and their records, to reconsider their display, both online and in-gallery, and to re-evaluate institutional responsibilities to a range of audiences (Taylor and Gibson, 2017). Actions such as the removal of certain items from display, the return of sacred objects to source communities, and the supplementing of records with additional contextual information about how materials came to be included in museum collections create a dialogue between the viewer and the object in question – the interaction with it is historicized and complicated, which in turn has profound implications for its use as a source of scholarship (Geismar, 2018). But as it stands, this process risks being truncated when the object is represented as RDF. While RDF provides a structure for representing the data, supplementary contextual information is modeled using other semantic tools, such as ontologies, controlled vocabularies, or authority files. Without ensuring that this information is also readily accessible to a querying tool, we risk seeing museum objects which have been incorporated into the Linked Data ecosystem as uniform and reducible to the conceptual logic of code, which ultimately entails a reduction in their complexity.

The ethnographic nature of materials in many museum collections presents a particular set of difficulties when it comes to standardizing data using controlled vocabularies or thesauri. These challenges can be broken into two distinct types: those related to the terminologies used when the records were created, and those related to the information content of the records themselves. The first set of challenges is one which is emerging as curators, archivists, and librarians have begun to approach their institutional practices and inherited records from reflective critical positions. Over the last decade

or so, scholarship on libraries, archives, and museums has begun to address the historical cataloging and documentation strategies of these institutions. The naming and classification of people and objects through the application of what in the eighteenth and nineteenth centuries were considered to be scientific methods has shaped the field of knowledge in museums and persists in many of the records today. While there is a growing awareness among museum professionals that they need to address, acknowledge, and ameliorate these histories, particularly as they are manifested in their records, there is also a question of stewardship. On the one hand, these data collections are the ideal context in which to approach these questions, since ethnographic museum data meets several of the technical and conceptual requirements for an in-depth inquiry: it is complex and heterogeneous, often irregular, and many different metadata schemas are in use in the field. Museum data also requires context in order to be understood and reused in ways which are sensitive to the origins and history of much of the material, which adds a level of urgency to the process. But on the other hand, scholars working on the creation of these links should also ask themselves the question of whether everything that *can* be linked *should* be. The collections data in many of these institutions is complicated and not always comfortable, but hiding this data is not a constructive response. As scholarship around responsible data stewardship develops (Coleman, 2020), it becomes imperative that we consider this tension, particularly as large cultural heritage institutions in Europe continue to open their collections. This is not an argument against the use of RDF as a technical representation format. Rather, it is a question of what models to use, so that we are able to preserve the complexity and ambiguity that make museums important sites of study, while not impeding interoperability. Models for both data and platforms which replicate the siloed nature of earlier data structures risk rendering Linked Data unusable, negating its purpose and squandering its key benefits.[16]

This state of affairs also raises the problem of how to include attribution in an RDF-based knowledge graph. In their study of the publication of medical data, Bechofer et al. argue that publishing scientific results as LOD poses the risk of confounding the flow of rights to the researcher. This has resonance with the use of RDF as a tool for representing and publishing some types of cultural heritage data. In this context, the question of rights is particularly significant, since copyright regimes and provenance metadata for cultural heritage collections is often extremely complex, as outlined in Section 4.1. The issues of attribution and contextualizing data are also significant for researchers and institutions who work on the provenance of looted and

[16]https://www.w3.org/standards/semanticweb/data

96

stolen artworks. This community was quick to realize the value of LOD as a tool for managing the documentation trails which are essential for tracking lost or stolen artworks, and, if necessary, establishing legal claims for ownership and provenance (Fink et al., 2014). While the linking of collections has increased the ability of researchers to follow certain individuals involved in looted or stolen art markets, there is little, at present, in the way of a centralized authority, like the gazetteers used by Recogito, against which to resolve their identities. Collaborative community initiatives such as Open Art Data are using Linked Data tools to create initial searches across collections, and highlight or 'red flag' names of collectors and art dealers who were associated with the theft of artworks. However, provenance tracking in these situations is extremely complex, as false provenance information is regularly being inserted into the documentation of looted and stolen works.[17] In these cases, without the addition of contextualizing information, the creation of semantic links between collections may have the undesired effect of perpetuating certain falsehoods around the ownership and/or origin of objects, which poses the risk of confounding the ability of researchers to conduct digital network analysis of the markets for looted and stolen art.

7 Summary and Conclusion

In this paper, we have tried to assess the use of Linked Data from cultural heritage sources from both a practical and a theoretical perspective, and to show how data consumers are often confronted with a situation where they have both too much available data and not enough useful data. We have shown how Linked Data can provide a meaningful mechanism for creating connections between sources and for facilitating scholarly practices. But such an approach is only possible if there is a critical mass of data available, and if the connections between sources can be established through reliable, open, persistent authorities, such as gazetteers. Without these verification mechanisms, the mass of heritage LOD quickly becomes unnavigable and unusable, resulting in a famine of usable data. We have also looked at the practical difficulties of managing and integrating contextualizing metadata, such as provenance information, alongside RDF triples, and have argued that this can be a significant problem for certain heritage data types, such as ethnographic collections, or looted and stolen artworks. One of the implications of such difficulties is that data which is incorrect, culturally insensitive, or contains outdated terminologies can inadvertently be released into the LOD ecosystem, without a tempering mechanism to draw users' attention to the issues at hand. This state of affairs is particularly troubling for large collec-

---

[17]https://www.openartdata.org/2020/07/how-to-track-falsification-of-provenance.html

tions of heritage data, which are proliferating as more and more institutions open up their collections. At the same time, this proliferation of data belies the significant effort and complexity faced by museums when they undertake the process of converting, mapping, and verifying their data into a LOD schema. These difficulties are particularly pronounced in ethnographic collections, where the current turn towards reevaluating collections and their colonial past requires augmentations to the collections' documentation and, consequently, the available data. We have also shown how content licensing that facilitates open sharing, one of the essential components of Linked Data, can be imprecise and difficult to manage, which creates a risk of loss of data within the ecosystem.

References

Bechhofer, S., Buchan, I., De Roure, D., Missier, P., et al. (2013). Why Linked Data is Not Enough for Scientists. *Future Generation Computer Systems*, 29(2):599–611, DOI: `10.1016/j.future.2011.08.004`.

Beltrame, T. (2016). Creating New Connections: Objects, People and Digital Data at the Musée du Quai Branly. *Anuac*, 4(2):106–129, DOI: `10.7340/anuac2239-625X-1980`.

Berners-Lee, T. (1999). Realising the Full Potential of the Web. *Technical Communication*, 46(1):79–82, `https://www.learntechlib.org/p/85551`.

Blagoev, B., Felten, S., and Kahn, R. (2018). The Career of a Catalogue: Organizational Memory, Materiality and the Dual Nature of the Past at the British Museum (1970-Today). *Organization Studies*, 39(12):1757–1783, DOI: `10.1177%2F01708406618789189`.

Blanke, T. and Hedges, M. (2013). Scholarly Primitives: Building Institutional Infrastructure for Humanities E-Science. *Future Generation Computer Systems*, 29(2):654–661, DOI: `10.1016/j.future.2011.06.006`.

Blijden, J. (2018). The Accuracy of Rights Statements on Europeana.eu. Technical report, Kennisland.

Coleman, N. (2020). Managing Bias When Library Collections Become Data. *International Journal of Librarianship*, 1(4):8–19, DOI: `10.23974/ijol.2020.vol5.1.162`.

Davis, K. (2019). Old Metadata in a New World: Standardizing the Getty Provenance Index for Linked Data. *Art Libraries Journal*, 44(4):162–166, DOI: `10.1017/alj.2019.24`.

Fink, E., Szekely, P., and Knoblock, C. (2014). How Linked Open Data Can Help in Locating Stolen or Looted Cultural Property. In Ioannides, M., Magnenat-Thalmann, N., Fink, E., Žarnić, R., et al., editors, *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection*, pages 228–237. Springer International Publishing, Cham, DOI: 10.1007/978-3-319-13695-0_2 2.

Geddes, M. (2019). Strategies to Support Wider Adoption of Linked Open Data in Smaller Museums. http://jhir.library.jhu.edu/handle/1774.2/62123.

Geismar, H. (2018). *Museum Object Lessons for the Digital Age*. UCL Press, London.

Gilliland, A. (2016). Setting the Stage. In Baca, M., editor, *Introduction to Metadata*. Getty Publications, https://www.getty.edu/publications/intrometadata/setting-the-stage/.

Griffiths, A. (2010). Collections Online: The Experience of the British Museum. *Master Drawings*, 48(3):356–367, https://www.jstor.org/stable/25767237.

Isaksen, L., Simon, R., Barker, E. T., and de Soto Cañamares, P. (2014). Pelagios and the Emerging Graph of Ancient World Data. In *Proceedings of the 2014 ACM Conference on Web Science*, WebSci '14, pages 197–201, New York, NY. Association for Computing Machinery, DOI: 10.1145/2615569.2615693.

Knoblock, C. A., Szekely, P., Fink, E., Degler, D., et al. (2017). Lessons Learned in Building Linked Data for the American Art Collaborative. In d'Amato, C., Fernandez, M., Tamma, V., Lecue, F., et al., editors, *The Semantic Web – ISWC 2017*, pages 263–279. Springer International, DOI: 10.1007/978-3-319-68204-4_2 6.

McDonough, K. and van der Camp, M. (2017). Mapping the Encyclopédie: Working Towards an Early Modern Digital Gazetteer. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities - GeoHumanities'17*, pages 16–22. ACM Press, DOI: 10.1145/3149858.3149861.

Murrieta-Flores, P. and Gregory, I. (2015). Further Frontiers in GIS: Extending Spatial Analysis to Textual Sources in Archaeology. *Open Archaeology*, 1(1):166–175, DOI: 10.1515/opar-2015-0010.

Sikos, L. F. and Philip, D. (2020). Provenance-Aware Knowledge Representation: A Survey of Data Models and Contextualized Knowledge Graphs. *Data Science and Engineering*, 5:293–316, DOI: 10.1007/s41019-020-00118-0.

Simon, R., Vitale, V., Kahn, R., Barker, E., et al. (2019). Revisiting Linking Early Geospatial Documents with Recogito. *e-Perimetron*, 14(3):150–163.

Taylor, J. and Gibson, L. K. (2017). Digitisation, Digital Interaction And Social Media: Embedded Barriers to Democratic Heritage. *International Journal of Heritage Studies*, 23(5):408–420, DOI: 10.1080/13527258.2016.1171245.

Turner, H. (2015). Decolonizing Ethnographic Documentation: A Critical History of the Early Museum Catalogs at the Smithsonian's National Museum of Natural History. *Cataloging & Classification Quarterly*, 53(5/6):658–676, DOI: 10.1080/01639374.2015.1010112.

Wallace, A. and Euler, E. (2020). Revisiting Access to Cultural Heritage in the Public Domain: EU and International Developments. *IIC - International Review of Intellectual Property and Competition Law*, 51:823–855, DOI: 10.2139/ssrn.3575772.