2018

Linked Data for Archivists: Graphs and Rhizomes

Matthew Roland Miguez

Linked Data for Archivists: Graphs and Rhizomes

Matthew Miguez

This article is adapted from a session presented at the joint meeting of the Society of Florida Archivists and Society of Georgia Archivists in Savannah, Georgia in October 2016.

The data models used in libraries and archives are in the midst of a big change. Traditional bibliographic records use the structure of relational databases for information storage and retrieval. Adoption of linked data technologies and supporting practices—such as Bibliographic Framework (BIBFRAME) and Resource Description and Access (RDA)—illustrate the movement from a relational data model to a graph-based model. Linked data is the underlying architecture of the semantic web, but many archives have attempted only tentative explorations of the technology. Linked data presents archivists with both challenges and opportunities by allowing archives to expand the scope of their descriptive practices, offering different types of data sources, and providing different voices a role in resource description. Conversely, linked data requires new skills and challenges the theoretical model of hierarchical arrangement by presenting a new data model: the rhizome, a structure prizing connection and relationship over arrangement. Archivists have a responsibility to be involved in the overall endeavor for their own benefit as well as for the benefit of others. The uniqueness of archival collections is a great benefit in expanding the data in the semantic web and linked data environment, and linked data can provide users of archives with a more fluid and complete discovery experience.

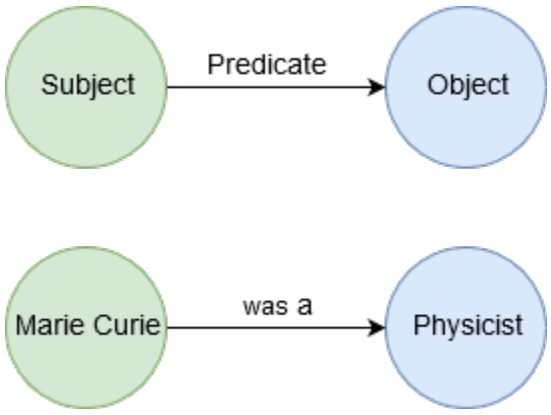
There is growing awareness in archival literature that the creation and organization of records are not benign acts, but, rather, operations of political interpretation and memory.[1] One problematized practice is archival description. In library literature, the conversation of cataloger bias dates to the 1970s[2]. Melanie Feinberg summarizes that, conscious or not, the biases and experiences of the cataloger enter the tools and practices of description and influence how resources are represented and used.[3] The inherently political nature of records compounds the challenge of describing archival materials,[4] as the representation of records can change the

[1]These ideas are most thoroughly explored in *Archival Science* 2, issues 1-4 (2002).
[2]Sanford Berman, *Prejudices and Antipathies: A Tract on the LC Subject Headings Concerning People* (Jefferson, NC: McFarland & Co., 1993).
[3]Melanie Feinberg, "Hidden Bias to Responsible Bias: An Approach to Information Systems Based on Haraway's Situated Knowledges," *Information Research* 12, no. 4 (October 2007). http://www.informationr.net/ir/12-4/colis/colis07.html.
[4]John Ridener, *From Polders to Postmodernism: A Concise History of Archival Theory* (Duluth, Minn.: Litwin Books, 2009).

interpretation of their real-world consequences and affects.[5] Privileging certain aspects of material representation and access is even encoded in the guidelines for arrangement and description: "Each mission will lead to the high prioritization of certain users, so that access tools developed will address their particular informational needs. This will affect the extent of descriptive work and the types of products developed by arrangement and description."[6] A single voice is insufficient to describe the totality of what a record or a collection represents. More diverse and inclusive models, practices, and philosophies are necessary to serve as a counterweight to these limitations.

Power and politics have also problematized archival arrangement. In the 1980s, Gilles Deleuze and Félix Guattari identified "arborescent," or top-down hierarchies, as structures encoding and concentrating power in the status quo[7]; yet, 120 years since its conception, hierarchical organization is still the dominant philosophical model for archival arrangement. In opposition to the arborescent organizing principle, Deleuze and Guattari conceived of the rhizome—a structure of organization where the defining feature is connectedness and every part connects to every other part. There are many paths into, through, and out of the rhizome, thereby replacing boundaries and form with integration and comprehensiveness. In the context of information search and retrieval, a user would not necessarily be limited to the "authorized" arrangement and representation of resources defined by the archivist. Records in a rhizomatic data structure could connect organically across *fonds* and institutional borders. Data siloed in disparate discovery systems could integrate into more holistic narratives with the addition of greater context and more transparent connection between collections. Thus, the rhizome, as conceived by Deleuze and Guattari, equally describes the graph-model underlying linked data.

Beginning in 1998, Tim Berners-Lee laid out the framework for linked data and the semantic web on his web site *Design Issues*.[8]



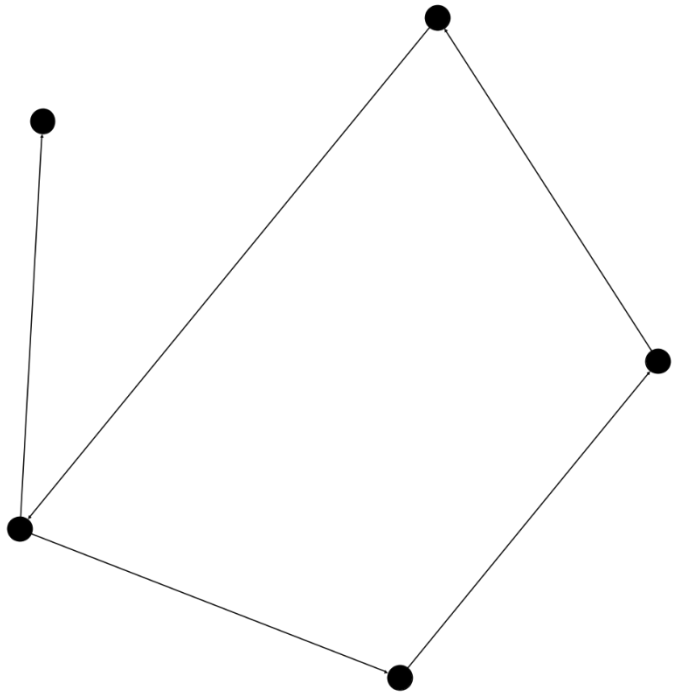
Model of an RDF triple.

The World Wide Web Consortium subsequently designed the Resource Description Framework (RDF) as a practical application of his graph-based metadata model. In RDF, tabular or

[5]Stacy Wood, Kathy Carbone, Marika Cifor, Anne Gilliland, and Ricardo Punzalan, "Mobilizing Records: Re-Framing Archival Description to Support Human Rights," *Archival Science* 14, no.3-4 (October 2014): 397-419.
[6]Kathleen Roe, *Arranging & Describing Archives & Manuscripts*, Archival Fundamentals Series: II (Chicago: Society of American Archivists, 2005).
[7]Gilles Deleuze and Félix Guattari, *A Thousand Plateaus: Capitalism and Schizophrenia* (London: Athlone, 1988).
[8]Tim Berners-Lee, "Design Issues for the World Wide Web," *Design Issues*, accessed February 28, 2018, https://www.w3.org/DesignIssues/.

relational data is atomized into three-part statements or triples: subject, predicate, and object.[9] These three-part statements allow metadata to be comprehensible in contexts apart from the original. The addition of unique identifiers (URIs), allow metadata to be linked to other data sources. Much like HTML web pages, anyone can publish RDF. Data described and linked through RDF create a mathematical structure called a graph. The scope of an RDF graph can be changed by including and excluding data from a variety of different sources.

In the fields of information science, graphs are gaining attention as a data model.[10] A graph is composed of two elements:



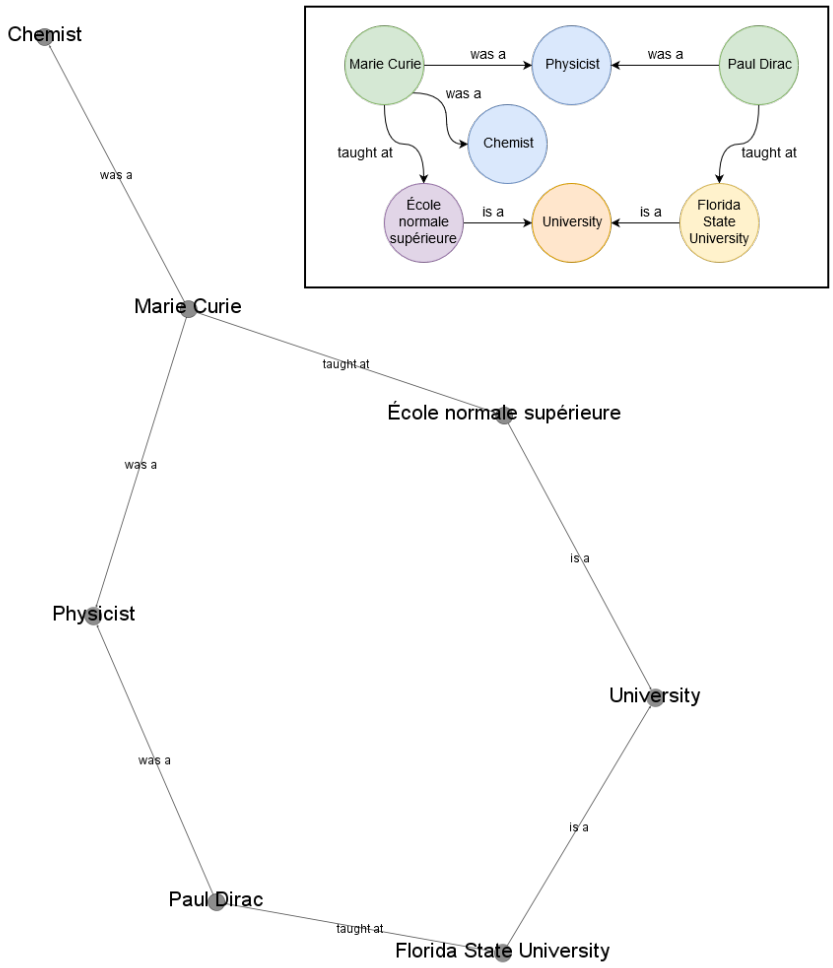
A simple graph of five nodes and five edges.

- Nodes: points on a plane
- Edges: lines connecting nodes

Google's highly successful PageRank search algorithm is based on treating the parts of the web it indexes as a graph;[11] each web page is a node, and hyperlinks are the edges connecting nodes. Mathematical measures of nodes in the graph then help Google construct the results delivered to the user. Graphs provide insights in other contexts as well. For example, digital humanists construct graphs of social networks. Moreover, traffic patterns can be analyzed by engineers using graphs. Furthermore, citation networks are another example of a graph in practice. In the context of resource description, each atomized datum is represented as a node in the graph. Edges represent relationships or links between those data points. RDF's subject, predicate,

[9]Dean Allemang and James A. Hendler, *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL* (Waltham, MA: Morgan Kaufmann/Elsevier, 2011).
[10]Ronald J. Murray and Barbara B. Tillett, "Cataloguing Theory in Search of Graph Theory and Other Ivory Towers," *Information Technology & Libraries* 30, no.4 (December 2011): 172.
[11]James E. Powell, Daniel Alcanzar, Matthew Hopkins, Robert Oldendorf, Tamara M. McMahon, Amber Wu, and Linn Collins, "Graphs in Libraries: A Primer," *Information Technology & Libraries* 30, no. 4 (December 2011): 158.

object language encodes such links: subjects and objects are nodes, connected through a relationship described by the predicate.



Seven RDF triples form a graph.

Thus linked data is a technological implementation of information as a graph, and a graph, with its flexibly structured but connected data, is an implementation of a rhizome.

Berners-Lee's vision is for linked data to do for data what the World Wide Web technologies did for documents. This semantic web allows for greater participation with data. If RDF follows the same path as HTML, tools and systems will be developed making the creation and publication of linked data easy. Marginalized and underrepresented communities will be able to engage in the description and use of resources and information, allowing their views and values to supplement description created by professional catalogers and archivists. There is potential for the inherent bias in top-down description to be mitigated through greater descriptive participation enabled by linked data. More voices can lead to greater discovery, as patrons with varied experiences and domain knowledge will have a broader descriptive landscape to query. Of course, this could also lead to conflicting description coexisting within the same information landscape. Archives can limit their discovery systems to query only "verified" or "authorized" linked data sources, but there is nothing stopping a linked data

developer or "power-user" consumer of linked data from incorporating linked data published by an archive into a search system with their own sources of description. This can lead to more varied use of resources. As described earlier, archives can only anticipate a limited number of uses and describe their collections accordingly. As more voices are involved, the use and description of resources can increase dramatically as historically marginalized groups find themselves more equitably represented in the description of archival materials.

Linked data has the potential to aid in serendipitous discovery, and archives and special collections have the most to gain in this area.[12] Since access to materials is typically mediated and supervised, remote users are limited to the information they can cull through the archives' discovery systems. Linked data and its associated query technology (SPARQL)—along with a search interface that includes linked data namespacing—will allow a user to tailor a discovery environment that meets their personal information needs. Traditional library OPACs present users with a web-based search interface that queries a single bibliographic data store. A linked data catalog built using the same model can include the option of querying not just the local RDF triple store, but other linked data graphs distributed across a wide range of linked data publishing platforms. Data about materials can be collocated from a variety of linked sources, enabling searches across several institutions or even knowledge domains. The ability to navigate beyond the strictures of provenance and institutional boundaries could enable a more thorough information experience. Hidden connections between collections and materials would be more readily apparent. Data silos can be minimized, and our users will be empowered to make surprising discoveries and interesting connections between collections.

Archives are uniquely situated to be productive participants in extending the graph of the semantic web. While there has been a more robust exploration of linked data on the library side of the information sciences, their holdings are largely non-unique and overlapping. Additionally, local cataloging practices tend to put limits on the amount of time and description applied to each resource. The amount of information libraries can add to the linked data graph is naturally limited. Though vast, the number of published resources are finite and libraries' collections often duplicate and overlap one another. Archives, on the other hand, have a wealth of unique collections and in the past revisited and refined description practices to fit new sources of information or changing user needs.[13] The contributions of archivists could grow and continually update and refine the linked data graph.

Archives also receive many benefits from publishing and consuming linked data. Archival description built on top of linked data triples can be enhanced by including triples published by other groups or institutions.[14] This more robust description would be a boon to local discovery, guiding users to resources from external data sources. Additionally, external discovery is positively impacted when description of another institution's collections is connected and linked with the description of collections at the local institution.

Currently, the production of linked data is difficult because tools tend to be technical and application specific.[15] Most institutions creating linked data do so by converting existing data,[16] but EAD is not a great source candidate for conversion into RDF. Karen F. Gracy notes, "EAD privileges the narrative character of the finding aid"[17] over the atomized access points required

[12]Getaneh Alemu, Brett Stevens, Penny Ross, and Jane Chandler, "Linked Data for Libraries: Benefits of a Conceptual Shift from Library-Specific Record Structures to RDF-Based Data Models," *New Library World* 113, no. 11/12 (2012): 549-571.
[13]Though this is less common post-MPLP.
[14]Enriching those records created by following MPLP.
[15]Alemu et al., 557.
[16]Jinfang Niu, "Linked Data for Archives," *Archivaria* 82 (Fall 2016): 83-110.
[17]Karen F. Gracy, "Archival Description and Linked Data: A Preliminary Study of Opportunities and Implementation Challenge," *Archival Science* 15, no. 3 (September 2015): 249.

by RDF. Some modern digital asset management systems use RDF internally, and there is no technological reason why all archival management software cannot create and publish RDF automatically and in parallel to the routine activities of archival arrangement and description. As producing and consuming linked data becomes a more common practice, better tools and more accessible workflows will be developed.

In the meantime, there are steps archivists and archives can take to plan for future linked data creation. If an archive has technological resources, they can ask developers to begin exploring RDF and related technologies. This will ensure the institution is prepared when publishing and consuming RDF becomes more feasible. Within the community, archivists can ask for guidance and exploration of linked data solutions from peers and professional associations. Encoded Archival Description (EAD) has only limited support for key features of linked data such as URIs, and that same support in EAD3 is merely experimental.[18] Collective interest can drive future revisions of EAD to be more amenable to linked data uses. Finally, as institutions look towards transitioning to EAD3, workflows can be examined to see if some linked data features can be incorporated. The University of Florida Libraries (UF) recently updated those XSLT stylesheets that generate HTML from EAD to include RDFa attributes in their HTML finding aids.[19] RDFa is a method of embedding RDF triples in HTML. Now, users with the corresponding technical skills can scrape UF's HTML finding aids to harvest the triples embedded within and develop a graph of the UF Libraries' collections. Relatively minor actions such as these can prepare institutions and the profession for the next development in information discovery and retrieval.

Development of the semantic web is an ongoing process. The library and archives communities have much to offer the overall endeavor, but limited resources and expertise hold them back from participating fully. Archival institutions—which have the most to contribute and the most to gain—have made only tentative advances in exploring linked data. Linked data provides archivists an opportunity to rethink how they record and interact with archival description. The data model underlying linked data and the semantic web is very different from the top-down hierarchical model familiar in archival practice. A rhizomatic data model offers many benefits as well as challenges, and, in many cases, will involve archivists offering intellectual control to others. This prospect might be frightening, but it brings with it a fuller, richer, and more dynamic information environment.

[18]Kelcy Shepherd, "Using EAD3," in *Putting Description Standards to Work*, ed. Kris Kiesling and Christopher J. Prom (Chicago: Society of American Archivists, 2017): 178.
[19]Allison O'Dell and Matthew Miguez, "Linked Data for Archives and Easy Steps to Linked Archival Metadata" (presentation, Society of Florida Archivists Annual Meeting, Savannah, GA, October 14, 2016).

References

Alemu, Getaneh, Brett Stevens, Penny Ross, and Jane Chandler. "Linked Data for Libraries: Benefits of a Conceptual Shift from Library-Specific Record Structures to RDF-Based Data Models." *New Library World* 113, no. 11/12 (2012): 549-71.

Allemang, Dean, and James A. Hendler. *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*. Waltham, MA: Morgan Kaufmann/Elsevier, 2011.

Berman, Sanford. *Prejudices and Antipathies: A Tract on the LC Subject Heads Concerning People*. Jefferson, N.C.: McFarland & Co., 1993.

Berners-Lee, Tim. "Design Issues for the World Wide Web." *Design Issues*. Accessed February 28, 2018. https://www.w3.org/DesignIssues/.

Deleuze, Gilles, and Félix Guattari. *A Thousand Plateaus: Capitalism and Schizophrenia*. London: Athlone, 1988.

Feinberg, Melanie. "Hidden Bias to Responsible Bias: An Approach to Information Systems Based on Haraway's Situated Knowledges." *Information Research* 12, no. 4 (October 2007). http://www.informationr.net/ir/12-4/colis/colis07.html.

Gracy, Karen F. "Archival Description and Linked Data: A Preliminary Study of Opportunities and Implementation Challenges." *Archival Science* 15, no. 3 (September 2015): 239–94. https://doi.org/10.1007/s10502-014-9216-2.

Murray, Ronald J., and Barbara B. Tillett. "Cataloging Theory in Search of Graph Theory and Other Ivory Towers." *Information Technology & Libraries* 30, no. 4 (December 2011): 170–84.

Niu, Jinfang. "Linked Data for Archives." *Archivaria* 82 (Fall 2016): 83–110.

O'Dell, Allison, and Matthew Miguez. "Linked Data for Archives and Easy Steps to Linked Archival Metadata." Presented at the Society of Georgia Archivists and Society of Florida Archivists 2016 Joint Annual Meeting, Savannah, GA, October 14, 2016.

Powell, James E., Daniel Alcazar, Matthew Hopkins, Robert Olendorf, Tamara M. McMahon, Amber Wu, and Linn Collins. "Graphs in Libraries: A Primer." *Information Technology & Libraries* 30, no. 4 (December 2011): 157–69.

Ridener, John. *From Polders to Postmodernism: A Concise History of Archival Theory*. Duluth, Minn.: Litwin Books, 2009.

Roe, Kathleen. *Arranging & Describing Archives & Manuscripts*. Archival Fundamentals Series: II. Chicago: Society of American Archivists, 2005.

Shepherd, Kelcy. "Using EAD3." In *Putting Descriptive Standards to Work*, ed. Kris Kiesling and Christopher J. Prom (Chicago: Society of American Archivists, 2017), 158–238. Trends in Archives Practice Module #18 of the Trends in Archives Practice Series.

Wood, Stacy, Kathy Carbone, Marika Cifor, Anne Gilliland, and Ricardo Punzalan. "Mobilizing Records: Re-Framing Archival Description to Support Human Rights." *Archival Science* 14, no. 3–4 (October 2014): 397–419. https://doi.org/10.1007/s10502-014-9233-1