Big Data Analytics for Research Libraries and Archives

Timothy A. Thompson
Indiana University Bloomington
School of Informatics, Computing, and Engineering
Bloomington, Indiana 47408
timathom@indiana.edu

ABSTRACT

Research libraries and archives have played a longstanding role in information management and access. In the second half of the twentieth century, libraries were at the forefront of automation and networked access to information. Since the advent of the internet, however, they have failed to keep pace with technological advances and currently face serious challenges in serving the evolving needs of researchers, whose information-seeking strategies are now shaped by internet search engines and online social media applications. To remain relevant in the current information landscape, libraries and archives must implement new strategies for converting legacy metadata to new formats that can add value to the research process. Although the data and metadata produced by libraries and archives may not always qualify, *prima facie*, as big data, an awareness among information professionals of the tools, techniques, and affordances of big data can help make library services more relevant to researchers.

KEYWORDS

i523, HID340, Library Metadata, Archival Metadata, Linked Open Data, Data Conversion

1 INTRODUCTION

Cultural heritage institutions such as libraries and archives have a longstanding tradition of producing structured data—in the form of catalog records or finding aids—to describe their collections. In the twentieth century, library card catalogs were gradually replaced by machine-readable formats, the foremost of which were the MAchine Readable Cataloging (MARC) formats for bibliographic and authority data (standardized as ISO 2709 and ANSI/NISO Z39.2) [3].

Initial development of the core MARC format, commissioned by the Library of Congress, was finalized in 1968, when the first electronic catalog records were distributed [3]. Originally, MARC records were used to facilitate the automated creation of card catalogs, which remained the primary method of information retrieval in libraries until the 1980s, when online public access catalogs (OPACs) became available and the pace of automation began to accelerate [9]. It was not until 2004 that the MARC record format (stored as a binary file) was mapped to an XML schema, making it more amenable to computation and transformation [7]. Notwithstanding, the MARC format has now become increasingly archaic and hinders data sharing and interoperability between libraries and contemporary platforms for research and information retrieval, such as the World Wide Web.

2 IS LIBRARY METADATA BIG DATA?

Although libraries and other cultural heritage institutions have created millions of metadata records over time, even the largest catalogs fall short of the scale typically associated with big data. The entire catalog of the Library of Congress, an institution that holds over 13 million physical volumes, totals less than 100 gigabytes. By comparison, Twitter produces approximately 12 terabytes of data on a daily basis [2, p. 1527]. According to Teets and Goldner, "If you consider just the metadata representing the collection of printed and electronic works held by libraries, it really cannot be considered big data in its current meaning" [9]. However, if big data is defined more broadly as a set of methodologies for analysis and an ecosystem for data aggregation, then libraries clearly stand to benefit from adopting its tools and techniques.

In one view, big data constitutes a "social movement," shaped by alliances "among heterogeneous players in business, academia, and government" [2, p. 1527]. By undertaking projects focused on data modeling and mass conversion and migration of legacy data, libraries can position themselves to partner with other players and provide enhanced information retrieval services, exposing their metadata in contexts that are more relevant to the current needs of researchers. In addition, by adopting graph-based models that are native to the World Wide Web, libraries can merge their data more seamlessly with the wider universe of online data in order to "generate massive collections of new relationship assertions" [9].

By leveraging universal standards such as the Resource Description Framework (RDF), libraries, archives, and other cultural heritage institutions can uncover latent relationships that are currently buried in catalog records, connecting them to data from disparate sources and providing a "training set for all human knowledge" [9, p. 430]. Teets and Goldner suggest that the process of splitting catalog records into discrete, linkable statements could vastly expand the size and scope of library-created metadata: "From a single [record], we can extract relationships from co-authors, citations, geo-locations, dates, named entities, subject classification, institution affiliations, publishers and historical circulation information. From these relationships, we can connect to other works, people, patents, events, etc. Creating, processing and making available this graph of new assertions at scale is big data" [9, p. 431].

3 TOWARD BIG DATA

Both libraries and archives face particular challenges in attempting to embrace the ethos of big and complex data. The rules and instructions used by catalogers and archivists to describe information-bearing resources are still reflective of the card catalog environment and do not support the kind of data-centric granularity needed to enable effective data integration and interoperability [10]. One of

the primary obstacles in converting and migrating legacy data is the problem of entity resolution and name disambiguation. A second obstacle, one that is by turns social, legal, and technical in nature, involves libraries' ability to publish and preserve digitized content.

3.1 Entity Resolution

Two recent projects exemplify the large-scale effort in libraries and archives to remediate legacy data and merge information from multiple sources. In the archival community, researchers are often faced with scenarios in which a person's papers are scattered among geographically distant repositories, but there is no master index that links the relevant collections together. One initiative, the Social Networks and Archival Context Project (SNAC), is working to develop algorithms and routines for entity resolution in order to address this problem. Researchers in the SNAC Project have focused on developing supervised machine learning algorithms for matching names across records that have been collected from multiple archival repositories [6]. Experiments with methods based on Naive Bayes Classification have yielded promising results, particularly when data from name strings is combined with contextual information (such as birth and death dates) that has been extracted from related records, with an accuracy rate of approximately 80% [6].

The task of entity resolution is made particularly difficult by the approach to data creation that has been traditionally employed by libraries and archives. In catalog records describing a book or archival collection, for example, creators are identified by name strings rather than unique identifiers. Catalogers must follow detailed rules for ensuring that each name string—known as an "authorized heading"—is unique, but because these strings are handcrafted by humans rather than generated by machines, they are particularly vulnerable to error and inconsistency.

In the Wikidata database, by contrast, which was originally compiled from structured data templates on Wikipedia pages, the American author Mark Twain is represented by a unique identifier that can be dereferenced as an HTTP URI: https://www.wikidata.org/wiki/Q7245. Wikipedia pages about Twain, regardless of the language they are written in, are able to link to this single identifier. In the Library of Congress Name Authority File, however, Twain is identified instead by the string "Twain, Mark, 1835-1910."

The Library of Congress maintains the "authorized" list of names for U.S. libraries, but many other national libraries maintain their own authority files. In the case of a well-known author such as Twain, there may be substantial agreement across institutions from Roman-script language communities as to the format of the authorized heading. For libraries and archives whose official languages are expressed in other character sets, the process of entity resolution may be more difficult.

To address the problem of string-based identification, the OCLC Online Computer Library Center, a global data provider for the library industry, has developed an initiative called the Virtual International Authority File (VIAF). VIAF is a data aggregation portal that attempts to resolve named entities across "more than 130 million authority and bibliographic records expressed in multiple languages, scripts, and formats" [4, p. 1]. MARC authority records are contributed to VIAF from nearly 50 contributing partners, most of which are national libraries [4]. References to named entities are clustered and merged using a 300+ core Hadoop cluster in a monthly batch process that takes approximately "12 hours of cluster compute time to complete" [4, p. 2]. In a multistep algorithm, named entities in VIAF are progressively grouped into identity clusters, and pair-wise matching is performed between datasets from each institutional contributor. Because it is able to draw on a wider range of sources for disambiguation and entity resolution, VIAF has, to date, achieved a higher degree of accuracy in matching entities than has the SNAC Project, with success rates of over 90% [4].

3.2 HathiTrust

In the library domain, the project that perhaps comes closest to the scale and scope of "big data" is the HathiTrust Digital Library and the related HathiTrust Research Center. In large part, the HathiTrust initiative grew out of the response of major research libraries to the Google Books mass digitization enterprise [1, 8, 11]. Libraries were particularly concerned about the issues of long-term digital preservation and open access to research data. With the favorable settlement of a high-profile lawsuit brought by the Authors Guild and other plaintiffs against both Google and HathiTrust, the latter has moved ahead with research projects to publish curated datasets extracted from the full text of both public domain and in-copyright titles. HathiTrust is committed to providing "non-consumptive" access to its data, and it has developed an approach that provides access through "data capsules"; this approach gives researchers as much flexibility as possible while simultaneously protecting against the unlawful "leakage" of full-text content onto the open web [11].

Researchers are now able to perform data mining on an extracted features dataset that contains page-level data features from all of the nearly 14-million volumes in the HathiTrust corpus. Although this dataset is substantially larger than the largest catalog of library metadata, its current size of 4 terabytes is still comparatively small by big data standards [5]. Nonetheless, HathiTrust's methodological sophistication and principled approach to data use and access provide a model for other projects in the library domain to follow.

4 CONCLUSION

For libraries, archives, and other cultural heritage institutions, the most significant paradigm shift that could be attributed to the big data phenomenon is a new view of descriptive metadata *as* data in its own right. As libraries in particular move away from legacy formats and domain-specific idiosyncrasies, they will be better equipped to serve the evolving needs and interests of researchers, who may themselves be struggling to come to grips with the scale of data in the age of the internet. Once libraries gain a more sophisticated understanding of their own data models and formats, they will be better positioned to assist researchers in managing, storing, and sharing their data—which is likely to be much bigger than anything produced by libraries themselves.

ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski and the i523 teaching assistants for their support and suggestions in writing this paper.

REFERENCES

[1] H. Christenson. 2010. HathiTrust: A Research Library at Web Scale. *LRTS* 55, 2 (2010), 93–102.
[2] H. Ekbia, M. Mattioli, I. Kouper, G. Arave, A. Ghazinejad, T. Bowman, V. R. Suri, A. Tsou, S. Weingart, and C. R. Sugimoto. 2015. Big Data, Bigger Dilemmas: A Critical Review. *Journal of the Association for Information Science and Technology* 66, 8 (2015), 1523–1545.
[3] K. M. Ford. 2012. LC's Bibliographic Framework Initiative and the Attractiveness of Linked Data. *ISQ: Information Standards Quarterly* 24, 2/3 (2012), 46–50. http://www.niso.org/publications/isq/2012/v24no2-3/ford/
[4] T. B. Hickey and J. A. Toves. 2014. Managing Ambiguity in VIAF. *D-Lib Magazine* 20, 7/8 (2014), 1–12.
[5] A. Kinnaman and E. Dickson. 2017. HTRC Docs: Extracted Features Dataset. (Sept. 2017). https://wiki.htrc.illinois.edu/display/COM/Extracted+Features+Dataset accessed 2017.
[6] R. R. Larson and K. Janakiraman. 2011. Connecting Archival Collections: The Social Networks and Archival Context Project. In *Research and Advanced Technology for Digital Libraries, TPDL 2011*. 3–14. https://doi.org/10.1007/978-3-642-24469-8-3
[7] Library of Congress. 2004. MARC XML Design Considerations. (Dec. 2004). http://www.loc.gov/standards/marcxml/marcxml-design.html accessed 2017.
[8] B. Plale, R. McDonald, Y. Sun, I. Kouper, R. Cobine, J. S. Downie, B. Sandore Namachchivaya, and J. Unsworth. 2013. HathiTrust Research Center: Computational Access for Digital Humanities Research and Beyond. In *JCDL'13*.
[9] M. Teets and M. Goldner. 2013. Libraries' Role in Curating and Exposing Big Data. *Future Internet* 5 (2013), 429–438. https://doi.org/10.3390/fi5030429
[10] R. Tennant. 2002. MARC Must Die. (Oct. 2002). http://lj.libraryjournal.com/2002/10/ljarchives/marc-must-die/ accessed 2017.
[11] J. Zeng, G. Ruan, A. Crowell, A. Prakash, and B. Plale. 2014. Cloud Computing Data Capsules for Non-Consumptive Use of Texts. In *ScienceCloud 2014*.

5 BIBTEX ISSUES

Warning–empty publisher in rL11

Warning–empty address in rL11

Warning–empty publisher in bP13

Warning–empty address in bP13

Warning–page numbers missing in both pages and numpages fields in bP13

Warning–empty publisher in jZ14

Warning–empty address in jZ14

Warning–page numbers missing in both pages and numpages fields in jZ14

(There were 8 warnings)

6 ISSUES

DONE:

Example of done item: Once you fix an item, change TODO to DONE

6.1 Uncaught Bibliography Errors

Citations in text showing as [?]: this means either your report.bib is not up-to-date or there is a spelling error in the label of the item you want to cite, either in report.bib or in report.tex

6.2 Formatting

Incorrect number of keywords or HID and i523 not included in the keywords

Other formatting issues - missing references section

6.3 Character Errors

Pasting and copying from the Web often results in non-ASCII characters to be used in your text, please remove them and replace accordingly. This is the case for quotes, dashes and all the other special characters. - possibly the issue in report.bib