Library Support of Data-Driven Discovery @ Texas A&M: A Proposed Roadmap

Dr. Bruce Herbert | Director of the Office of Scholarly Communications
University Libraries | Texas A&M University
5000 TAMU | College Station, TX 77843-5000

Data-driven Discovery

Data-driven discovery is transforming the nature of research and scholarship across a broad swath of disciplines[1]. This evolution is possible because of advances in cyberinfrastructure, machine learning, and the availability of new and vast data streams. Data-driven discovery involves the meta-analysis of aggregated data from many experiments or studies (i.e. small data) or the analysis of very large data sets (i.e. big data) that arise from the aggregated behaviors of many agents or individuals. Data-driven discovery, whether focused on big or small data, requires the computer technology infrastructure to manage, store, and analyze the data sets; machine learning or data mining techniques for data analysis; and the metadata, ontological, or data model frameworks that allow heterogeneous data to be managed and analyzed. In addition, systems have to be developed or perfected that allow the data to be curated and preserved in useful formats while protecting privacy and security, as appropriate.

The library's role, as information specialists, in supporting data-driven discovery at Texas A&M is best focused on the development of structured, curated data sets that use ontologies and meta data schemas that can organize the highly heterogeneous data streams that make up big data or help aggregate small data. In addition, a library data expert can also have a significant role in the development of data curation and preservation systems. We propose recruiting a faculty with expertise in information science that is focused on partnering with research groups and centers that are using data-driven discovery to conduct innovative and translational research. This individual would be key members of research teams and would directly contribute to proposal development and funding.

The federal government and many private funding agencies now require the deposit of published articles that are the result of funded research. At the same time, there is increasing interest in the potential impact of the sharing and reuse of scholarly datasets deposited in repositories that are open and can be accessed by other researchers. Many funding agencies, including NSF, now require that applicants must submit a Data Management Plan (DMP) as part of the application process. Per the NSF web site: "Proposals that do not include a DMP will not be able to be submitted."[2]

A&M University Libraries Support of the Analytics of Big Data

The University Libraries has both current and emerging capabilities and resources to support research, student instruction, and outreach within the context of Big Data.

The Texas A&M University Libraries has both current and emerging capabilities and resources to support research, student instruction, and outreach.

Current

Library expertise in data management and curation. The variety of big data, in terms of incompatible data formats, non-aligned data structures, and inconsistent data semantics, represents a serious challenge in the management and analysis of the data. Library expertise in information metadata schemas and ontologies could be useful to big data researchers so that their data is discoverable, curated and preserved. The library is developing our capacity to partner with campus researchers.

As an example, Sarah Potvin, Digital Scholarship librarian, is co-PI on a recently submitted NSF proposal submitted by the Biodiversity Research and Teaching Collections (BRTC). The working title of the proposal is "Beyond the

[1] Hey, T., Tansley, S., & Tolle, K., The Fourth Paradigm. Data-intensive Scientific Discovery: Redmond, WA, Microsoft Research.
[2] http://www.nsf.gov/eng/general/dmp.jsp

Voucher: Critical Database and Digital Asset Management System Improvements to Facilitate the Linking of Media, Field Catalogs, and Genetic Materials to the Voucher Specimens at Texas A&M University."

Library digital collections as big data. The digital collections we purchase as well as the Texas A&M repositories of TAMU scholarship are big data sets. The Libraries has been negotiating rights to text mining of the scholarly corpus we make available to the campus allowing researchers to analyze and visualize major conceptual and research strands published in scholarly literature.

In addition, TAMU document and data repositories are themselves a big data set that could be mined and analyzed to support research, teaching and outreach. For instance, OAKTrust, Texas A&M's institutional repository (http://oaktrust.library.tamu.edu), serves as the primary tool for the curation of Texas A&M's electronic theses and dissertations (ETDs) and is an emerging platform for other types of Texas A&M's scholarly publications. Our repository contains over 71,000 digital objects representing almost four terabytes of information that have been downloaded more than 36 million times in the last 7 years.

Emerging

Data/text mining of library digital collections. The library is developing the capacity to guide researchers interested in using sophisticated text mining algorithms to analyze a large corpus of documents. We recently have had a number of requests for support in performing these studies by faculty in a range of disciplines.

Data curation for TAMU research teams. As research collaborations grow, the tasks associated with managing and curated the artifacts developed by those collaborations become more demanding and imperative. Librarians can partner with these programs to develop IT systems, workflows and standards that support the discoverability, curation and use of team data and publications, as well as meet compliance with grant mandates. Systematic management and sharing of large team data sets and publications has shown to enhance research productivity and creativity, especially in teams engaged in interdisciplinary research[3].

Data analysis and visualization. The library is exploring expanding our capabilities to offer training and consultations in experimental design, data analysis, and visualization for both small and big data. We offer some of these services now through our class-based training and consultation in GIS.

Library Big Data Project: Data/Text Mining of Large Digital Collections of Scholarly Publications

The volume and variety of published scholarship represents both a boon and challenge for academic research. Research libraries, in concert with scholarly publishers, have responded to these challenges by converting academic publications from print to digital collections, thereby enhancing both access and discoverability of the research (Borgman, 2007). The increasing volume of published work, though, still challenges researchers to form syntheses of domain knowledge, especially when engaged in interdisciplinary research. Large digital collections of published scholarly represent one type of big data sets (March 2006). Academic libraries and other curators of large digital collections are experimenting with text mining and other data mining and visualization tools to explore, visualize and synthesize large corpus of text (Witten et al., 2004, Lok 2010).

The digital collections we purchase as well as the Texas A&M repositories of TAMU scholarship are big data sets. The Texas A&M Libraries has been negotiating rights to text mining of the scholarly corpus we make available to the campus allowing researchers to analyze and visualize major conceptual and research strands published in scholarly literature. In addition, TAMU document and data repositories are themselves a big data set that could be mined and analyzed to support research, teaching and outreach. For instance, OAKTrust, Texas A&M's institutional repository (http://oaktrust.library.tamu.edu), serves as the primary tool for the curation of Texas A&M's electronic theses and dissertations (ETDs) and is an emerging platform for other types of Texas A&M's scholarly publications. Our repository contains over 71,000 digital objects representing almost four terabytes of information that have been downloaded more than 36 million times in the last 7 years.

[3] http://www.nap.edu/catalog/19007/enhancing-the-effectiveness-of-team-science

The library seeks to develop the systems that allow researchers to use sophisticated text mining algorithms to analyze a large corpus of documents (Weiss et al. 2010). We recently have had a number of requests for support in performing these studies by faculty in a range of disciplines. Our project represents a collaboration between librarians, statisticians, and IT specialists in the library.

References

Borgman, C.L. (2007). Scholarship in the Digital Age: Information, Infrastructure, and the Internet. MIT Press.
Lok, C. (2010). Literature mining: Speed reading. Nature 463: 416-418.
March, J. G. (2006). "What Do You Do with a Million Books?" D-Lib Magazine 12(3).
Weiss, S. M., Indurkhya, N., Zhang, T., & Damerau, F. (2010). Text mining: predictive methods for analyzing unstructured information. Springer Science & Business Media.
Witten, I. H., et al. (2004). "Text mining in a digital library." International Journal on Digital Libraries 4(1): 56-59.