Information Services & Use 37 (2017) 319–323
DOI 10.3233/ISU-170853
IOS Press


# Advancing library cyberinfrastructure for big data sharing and reuse

Zhiwu Xie [a,*] and Edward A. Fox [b]

[a] *University Libraries, Virginia Polytechnic Institute and State University, 560 Drillfield Drive Blacksburg, VA 24061, USA*

[b] *Department of Computer Science, Virginia Polytechnic Institute and State University, 114 McBryde Hall, M/C 0106, Blacksburg, VA 24061, USA*

**Abstract.** Data-intensive science presents new opportunities as well as challenges to research libraries. The cyberinfrastructural challenge, although chiefly technological, also involves social-economic and human factors, therefore requires a deep understanding of what roles research libraries should play in the research lifecycle. This paper discusses the rationale and motivations behind a research project to investigate effective library big data cyberinfrastructure strategies.

Keywords: Data management, Big Data, cyberinstrastructure, data sharing, data reuse

## 1. Introduction

As a key component of the nation's knowledge infrastructure, libraries must continuously reinvent themselves with the emergence and the establishment of new discovery paradigms. The recent wave of data-intensive science has motivated many high-profile library big data projects, notably the ambitious plan to archive all tweets at the Library of Congress [13], the heterogeneous and geographically-replicated archival storage known as the Digital Preservation Network (DPN) [15], the data mining facility at the HathiTrust Research Center (HTRC) [12], and the metadata hubs developed at the Digital Public Library of America (DPLA) [4] and the SHARE initiative [14]. Many more are being developed or being planned.

The common theme of these library projects is to handle high volumes of data. Since the volume usually exceeds the currently deployed capacity of the typical library cyberinfrastructure (CI), we must have more storage, processing capacity, and network bandwidth, to name a few requirements. It would be prohibitively expensive to build local, proprietary capacities at each library. Fortunately a wide range of shared CI options exist, including, but not limited to: 1) Institutional high-performance computing (HPC), high-throughput computing (HTC) and storage facilities, e.g., Indiana University's Big Red II, Virginia Tech's BlueRidge, etc.; 2) National HPC, HTC, and storage facilities, most notably XSEDE resources [16]; 3) National research clouds such as Chameleon Cloud, CloudLab, Open Science Data Cloud, etc.; 4) Commercial clouds, such as Amazon Web Services (AWS), Rackspace, etc.

These shared resources, especially the commercial clouds, have drastically lowered the barrier to entry for the big data game. This has become especially evident in the Information Technology (IT) industry,

* Corresponding author. E-mail: zhiwuxie@vt.edu.

where even a small start-up can handle high volumes of data today. Indeed, most library big data projects we surveyed are built on such shared CI resources and do not require significant initial investment. It is therefore a fallacy to assume that only libraries with deep pockets are qualified to provide big data services.

However, operating library big data services on shared CI resources is far from turnkey. Although some general guidelines exist [3,7,9] it is not always clear what we can learn from the IT sector's success in developing big data services. Are big data services to be operated at a research library the same as or different from those provided by common commercial services? What are the key technical challenges? What are the key performance characteristics? What are the monetary and non-monetary (time, skill set, administrative, etc.) costs? Are there any cost patterns or correlations among the CI options? What are the knowledge and skill requirements for librarians? To answer these questions, we must seek better understanding of why and how research libraries develop big data services.

## 2. Big data, small science, and libraries

Big data used to be closely associated with big science, which in turn is characterized by big organizations and big budgets [2,5,6]. This is no longer true. The fast advance and ubiquitous availability of sensing technologies, the Web, and the Cloud have erased the data volume boundary separating big and small science. For example, the 1000 Genomes project [1] produced 200 TB of data from 2008 to 2012. The Sloan Digital Sky Survey [18] produced about 130 TB of raw and derivative data over eight years in phases I and II. In contrast, the sensors installed in Virginia Tech's Goodwin Hall alone can collect 200 TB per year at moderate sampling rate, and a handful of Amazon EC2 instances can gather as much web data in weeks. Neither the big organization nor the big budget is a must-have to conduct data-intensive research. The Goodwin Hall project was started by two faculty members and a small lab. Crawling and analyzing web data is so affordable today that even a student can initiate his or her own web analytics project.

The leveling of the big data playing field makes it possible for many more small science projects to take advantage of big data. Organization-wise, these projects usually emerge from the ground up, with user communities naturally forming, growing, and self-organizing around the data connected with their own needs, use cases, and perspectives. However, these project teams usually lack experience and expertise to effectively extract values from the large data sets, therefore opening up the opportunities for the research libraries to build and offer new value-added services.

## 3. Use and reuse driven big data management

Even if cost is not a major concern, a traditional digital library can hardly deliver services suitable for data-intensive research, especially for small science projects. For example, a download link is no longer sufficient to provide effective access to terabytes of research data, because randomly-moving data to a remote site may take too long, and therefore clog the reuse workflow. When the data volume grows larger, it also becomes more difficult to justify a dark archive or traditional preservation approaches, e.g., making multiple copies and storing them among geographically-distributed locations. The ingestion and dissemination processing may involve much time and processing overhead, challenging the conventional wisdom that data may need to be stored in a different format, layout, or logical unit from when they are in use. In contrast to the more immediate user needs to effectively access data, these traditional digital

library functions become relatively trivial. Moreover, libraries are increasingly expected to deliver not just the raw data but also knowledge extracted from them, e.g., running user-specified algorithms against preserved data, pushing customized information from a metadata hub, or analyzing web archives.

It is therefore imperative to manage data with the use and reuse-driven approach [17]. Using the DCC Curation Lifecycle Model terms [8] developed by the Digital Curation Centre (DCC) in the UK, the library big data service should focus more on facilitating data use and reuse instead of spreading the library resources evenly among storage, preservation, resource description, and various transformations, each for its own purpose. Facilitating data use and reuse should become the driving force behind other activities. For example, under this approach, the storage layout should be optimized towards more efficient reuse, with multi-copy preservation considered a side effect of the replication deployed to increase access bandwidth.

## 4. Library big data service patterns

There will be many different library big data services, and each may be operated on any of the shared cyberinfrastructure options listed in Section 1. However, without appropriate categorization, it would be non-trivial trying to mix and match them to achieve optimal performance. We therefore conduct an environmental scan to extract service patterns.

Conceptually, we can draw three distinct, although not mutually-exclusive, service patterns, schematically shown in Fig. 1: 1) The Bridge Pattern clearly separates the data storage and data processing in different facilities, and answers sporadic, on-demand, and sometimes user-specified computing needs by moving data from storage to processing nodes through the network link between them. The Digital Preservation Network (DPN) nodes [2], despite being primarily concerned with data storage, may be considered special cases of the Bridge Pattern. This is because the data ingestion, validation, periodic fixity checking, and refreshing are indeed on-demand data processing performed at compute nodes away from where the data is stored. 2) The Network Pattern, as exemplified by warcbase [11] – an open-source platform for managing web archives built on Hadoop and HBase – features a much tighter integration between data storage and processing. Typically involving a Hadoop cluster, this pattern uses a large number of interconnected nodes, each serving both as a storage and processing unit. These nodes replicate, balance, and co-optimize both storage and computing across the interconnections. The Network Pattern excels at MapReduce types of computation and can sustain high processing loads. However, the

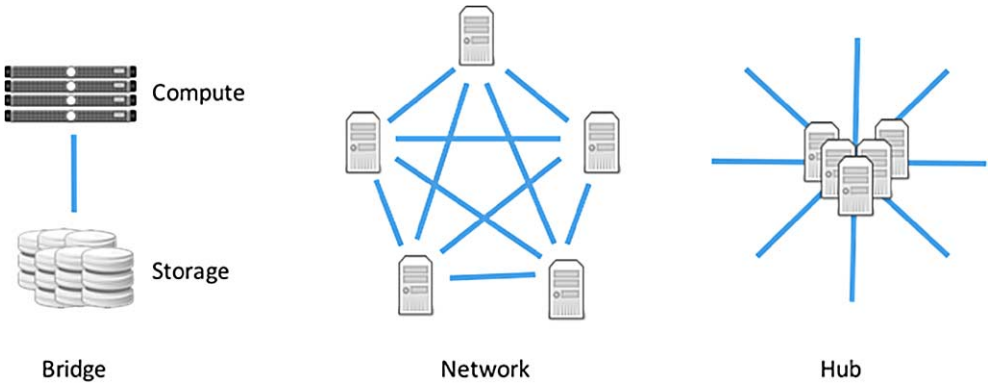


Fig. 1. Three patterns for library big data services.

initial data loading stage is known to be a bottleneck [10]. As a result, data tend to be "sticky" to the CI. Once loaded, the data usually stay put. 3) The Hub Pattern includes the Digital Public Library of America (DPLA) [4] and SHARE Notify [14], both specialized as metadata hubs. This service pattern continuously draws live data from potentially many sources, undertakes necessary processing, and then disseminates processed information to potentially large numbers of data consumers. It has higher quality of service (QoS) requirements, since downtime may lead to permanent data loss. In addition to the performance requirements on the computing and storage nodes, it also requires stable network connections to the external systems upon which it depends.

## 5. Summary

With an emphasis on big data sharing and reuse, we are conducting a research project aiming to develop an evidence-based, broadly-adaptable Cyberinfrastructure (CI) strategy to operate digital library services. The strategy will equip research libraries with knowledge and techniques to leverage shared CI resources and balance their desires, needs, and constraints with a clear understanding of the tradeoffs.

## Acknowledgement

This work was partially supported by IMLS LG-71-16-0037-16.

## About the authors

Zhiwu Xie is a professor and directs the digital library development team at Virginia Tech Libraries. He leads the development of the Goodwin Hall Living Lab data management system, IMLS ETDplus Workbench, VTechData, and UWS through transactional web archiving, among others. He served on technical committees of Fedora, APTrust, PREMIS, ResourceSync, and Altmetrics Data Quality efforts. His research extensively utilizes all types of CI options summarized in this paper, and has been supported by Mellon, IBM, Amazon, USGS, and NSF XSEDE.

Edward Fox is a professor in the Department of Computer Science, Virginia Tech. A Fellow of IEEE, he is a senior computer scientist and digital library innovator. He chaired the IEEE Technical Committee on Digital Libraries, ACM SIGIR, and the JCDL steering committee, and has been (co-)PI on one hundred and twenty-four research grants/contracts, (co-)authored eighteen books, one hundred and twenty-six journal/magazine articles, forty-nine book chapters, two hundred and nineteen refereed conference/workshop papers, seventy-four posters, and over one hundred and fifty other publications/reports, with an h-index of over fifty-six.

## References

[1] 1000 Genomes Project Consortium, A map of human genome variation from population-scale sequencing, *Nature* **467**(7319) (2010), 1061–1073. doi:10.1038/nature09534.

[2] J. Bicarregui, N. Gray, R. Henderson, R. Jones, S. Lambert and B. Matthews, Data management and preservation planning for big science, *International Journal of Digital Curation* **8**(1) (2013), 29–41. doi:10.2218/ijdc.v8i1.247.

[3] Cyberinfrastructure Council, *Cyberinfrastructure Vision for 21st Century Discovery*, National Science Foundation, Arlington, VA, 2007, Report No.: NSF-2007-28.

[4] Digital Public Library of America [Internet], [cited 2017 April 10]. Available from: https://dp.la/.
[5] N. Gray, T. Carozzi and G. Woan, Managing research data in big science, *JISC*, 2012 July, Available from http://arxiv.org/abs/1207.3923.
[6] P.B. Heidorn, Shedding light on the dark data in the long tail of science, *Library Trends* **57**(2) (2008), 280–299. doi:10.1353/lib.0.0036.
[7] G. Henry, Core infrastructure considerations for large digital libraries, Council on Library and Information Resources, Digital Library Federation, 2012.
[8] S. Higgins, The DCC curation lifecycle model, *International Journal of Digital Curation* **3**(1) (2008), 134–140. doi:10.2218/ijdc.v3i1.48.
[9] S.J. Jackson, P.N. Edwards, G.C. Bowker and C.P. Knobel, Understanding infrastructure: History, heuristics and cyberinfrastructure policy, *First Monday* **12**(6), 2007 June 4.
[10] Y. Kargın, M. Kersten, S. Manegold and H. Pirk, The DBMS – your big data sommelier, in: *2015 IEEE 31st International Conference on Data Engineering (ICDE)*, IEEE, 2015, pp. 1119–1130. doi:10.1109/ICDE.2015.7113361.
[11] J. Lin, M. Gholami and J. Rao, Infrastructure for supporting exploration and discovery in web archives, in: *Proceedings of the 23rd International Conference on World Wide Web*, ACM, 2014, pp. 851–856.
[12] B. Plale, R. McDonald, Y. Sun, I. Kouper, R. Cobine, J.S. Downie et al., *HathiTrust Research Center: Computational Access for Digital Humanities and Beyond*, Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, ACM, New York, NY, USA, 2013.
[13] M. Raymond, The Library and Twitter: An FAQ, | Library of Congress Blog [Internet], 2010 [cited 2017 April 10]. Available from http://blogs.loc.gov/loc/2010/04/the-library-and-twitter-an-faq/.
[14] SHARE Internet [cited 2017 April 10]. Available from http://www.share-research.org/.
[15] The Digital Preservation Network [Internet], [cited 2017 April 10]. Available from http://dpn.org/.
[16] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw et al., XSEDE: Accelerating scientific discovery, *Computing in Science & Engineering* **16**(5) (2014), 62–74. doi:10.1109/MCSE.2014.80.
[17] Z. Xie, Y. Chen, J. Speer, T. Walters, P.A. Tarazaga and M. Kasarda, Towards use and reuse driven big data management, in: *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, ACM, 2015, pp. 65–74. doi:10.1145/2756406.2756924.
[18] D.G. York, J. Adelman, J.E. Anderson Jr., S.F. Anderson, J. Annis, N.A. Bahcall et al., The sloan digital sky survey: Technical summary, *The Astronomical Journal* **120**(3) (2000), 1579. doi:10.1086/301513.