Editorial

Open Access

Weijia Xu*, Maria Esteva, Jessica Trelogan, Dan Wu

Special Issue on Cyberinfrastructure, Machine Learning, and Digital Library

https://doi.org/10.2478/dim-2019-0007

This special issue aims to bridge the gaps between digital libraries and archives and cyberinfrastructure by inviting researchers and practitioners from both fields as well as domain experts to share their ideas, introduce theories and methods, and demonstrate successful use cases. Increasingly, digital libraries and archives need to and are using cyberinfrastructure and machine learning to meet curation, data management, and researchers' needs. Academic libraries have made a significant progress accommodating data into their services and collections. This has been achieved through data management consulting services and institutional repositories for final and relatively small-sized data publications. However, research data management remains challenging for large-scale data generated from complex analysis pipelines conducted in distributed computational resources. More often than not, researches conducted in these ecosystems involve using multiple computational facilities, remote users with access and authorization roadblocks, data that evolve at varying time frames, missing documentation, data transfer problems, storage scalability limitations, and data vulnerability risks. A necessary approach to meeting these challenges is to use cyberinfrastructure, which refers to large shared online research environments, backed up by advanced computing resources hosted in data centers and supported by experts. Coupling cyberinfrastructure and digital libraries and archives can provide the needed technical resources and the expertise required to manage and analyze data at scale, as well as new opportunities to facilitate data preservation, access, and reuse.

Facilitating adoption and integration between machine learning, cyberinfrastructure and digital library is an important step to achieve smart data and information management. However, there is lack of mediums and forums that bring together researchers and practitioners to share visions, questions, latest advances in methodology, application experiences, and best practices. Library and archival professionals are often unfamiliar with cyberinfrastructure. In turn, cyberinfrastructure experts lack experience in traditional digital library and archives practices such as metadata, provenance, publishing, information retrieval, and digital preservation. To address this challenge, the workshop on cyberinfrastructure, machine learning, and digital library was held in conjunction with 2018 Joint Conference on Digital Libraries on June 3, 2018, in Dallas. The workshop included 10 project presentations and drew more than 40 participants at crossroad of these three fields. These projects span entire data life cycle and range from infrastructure perspective and service provider perspective to end-user perspective. From the workshop, five projects are invited to submit extended version for this special issue.

In the paper "Capsule Computing: Safe Open Science," Beth et al. presented their recent works and insights on enabling more options for data sharing. The authors argue that more flexible approaches are needed to share protected data and propose a solution named capsule framework. They demonstrated the synergies and tradeoffs of the current implementation through a use case of using their framework with a massive collection of copyrighted texts. Instead of just providing public data content and hiding any protected data for research, the proposed capsule framework enables a novel sociotechnical system involving a controlled interaction between humans, machines, and the environmental aspects of the work system. Through this framework, users can access, but with limited interaction, protected data. The limitations are imposed by the capsule framework to ensure that the data content is safe. HathiTrust data collection is used as an example data collection to demonstrate the usage and features of the proposed framework in the paper.

***Corresponding author: Weijia Xu,** Texas Advanced Computing Center, University of Texas at Austin, Austin, USA, Email: xwj@tacc.utexas.edu
Maria Esteva, Texas Advanced Computing Center, The University of Texas at Austin, Austin, TX, USA
Jessica Trelogan, University of Texas Library, TX, USA
Dan Wu, Center for Studies of Information Resources, Wuhan University, Wuhan, China

As data collection grows larger, data management, curation, and access require increasing computational capacity. Thomas et al. presented their experience and best practice for working with large data collection in their paper titled “Petabytes in Practice: Working with Collections as Data at Scale.” The paper reported progress using DRAS-TIC (digital repository at scale that invites computation), an open source archiving platform, and Brown Dog, a science driven data transformation service, as the basis for an expandable distributed storage/service architecture with on-demand, horizontally scalable integrated digital preservation and analysis services. The George Meany Memorial Archive has been used as a case study to demonstrate their approach. They also focused on technical approach of extracting text from image data and how this might be used to facilitate data access, such as using mobile device.

The trends of facilitating data access and adapting diverse platforms are also recognized in the paper “Predicting Academic Digital Library OPAC Users’ Cross-Device Transitions” by Liang and Wu. In this paper, they studied user’s cross-device transition behavior when using OPAC (online public access catalog). Based on the large-scale OPAC transaction log consisting of more than 16 million search query records over 6 months’ period, the online activities between device transitions in the process of using OPAC are used to predict user action after device transition. In the paper, they reported their methodology and findings in details as well as important features to consider. These results have potential to enable libraries and other data providers to provide smart services for users.

Another challenge for large digital collection is to uniquely track datasets through multiple copies, modifications and derivatives. This challenge is caused by both increasing complexity of dataset and broad data analysis methods available. Maria et al. reported Identifier Services (IDS) prototype for managing distributed genomics datasets remotely over time in “Identifier Services: Modeling and Implementing Distributed Data Management in Cyberinfrastructure.” The IDS enable tracking, comparison, and verification of same/similar dataset stored and distributed. In their paper, the architecture and workflow of IDS are discussed in a great detail. Six datasets are used for test and demonstration in the paper. Both computational results and user feedback are collected and reported in the paper. The results show that the prototype can effectively detect inconsistency between content and apparent data description and it is a welcome innovation for researchers of the data product.

Machine learning methods have been quickly adopted in many scientific fields as a new research approach. For digital libraries, machine learning methods can be used not only as research methods but also to enable additional services and features’ improvement throughout entire data life cycle. In the paper “*Improving Publication Pipeline with Automated Biological Entity Detection and Validation Service*”, an ensemble machine learning method is used to identify important terms from manuscripts in order to improve the readability and accessibility of scientific publications. A live web service has been implemented and integrated with publication pipeline used by American Society of Plant Biologists (ASPB). In addition to discussion of technical approaches and evaluations, the paper also describes service deployment and reports production status to show its practical impact.

Together, the five papers demonstrate how digital library, cyberinfrastructure, and machine learning can be integrated together to provide and enhance new features for data collection management and reuse. The five selected papers also cover different stages of data life cycle, from data ingest, data access, data analysis, data curation, and data archives. We hope this special issue can inspire interested readers and make more contributions to this dynamic field.