On the Computational Turn in Archives & Libraries and the Notion of Levels of Computational Services

Greg Jansen, Richard Marciano
Digital Curation Innovation Center (DCIC)
University of Maryland's iSchool

1. The Computational Turn in Archives & Libraries

The University of Maryland iSchool's Digital Curation Innovation Center (DCIC) is pursuing a strategic initiative to understand and contribute to the computational turn in archives and libraries. The foundational paper (with partners from UBC, KCL, TACC, and NARA) calls for re-envisioning training for MLIS students in the "Age of Big Data". See: **"Archival Records and Training in the Age of Big Data"** at http://dcicblog.umd.edu/cas/about/. We argue for a new Computational Archival Science (CAS) inter-discipline, with motivating case studies on:

1. evolutionary prototyping and computational linguistics,
2. graph analytics, digital humanities and archival representation,
3. computational finding aids,
4. digital curation,
5. public engagement / interaction with archival content,
6. authenticity, and
7. confluences between archival theory and computational practices: cyberinfrastructure and the records continuum.

Deeper experimentation with these new cultural computational approaches is urgently needed and the DCIC is developing a CAS curriculum that brings together faculty from Computer Science, Archival & Library Science, and Data Science. We conduct experiential projects teams of students to help them: gain digital skills, conduct interdisciplinary research, and explore professional development opportunities at the intersection of archives, big data, and analytics. These projects leverage unique types of archival collections: refugee narratives, community displacement, racial zoning, movement of people, citizen internment, and cyberinfrastructure for digital curation. See **"Practical Digital Curation Skills for Archivists in the 21st Century"** (Lee, Kendig, Marciano, Jansen), MARAC 2016, at: http://bit.ly/2lL5Et9. **Two workshops** on the interplay of computational and archival thinking were held in:

- April 2016 (http://dcicblog.umd.edu/cas/), and
- December 2016 (http://dcicblog.umd.edu/cas/ieee_big_data_2016_cas-workshop/)

and a **pop-up session at SAA 2016** discussed archival records in the age of big data: https://archives2016.sched.com/event/7f9D

Finally, the DCIC is developing new cyberinfrastructure, called ***DRAS-TIC*** (see Nov. 2016 CNI talk at: https://www.cni.org/topics/digital-curation/drastic-measures-digital-repository-at-scale-that-invites-computation-to-improve-collections), that facilitates computational treatment of

cultural data. *DRAS-TIC* stands for Digital Repository at Scale that Invites Computation (To Improve Collections), and blends hierarchical archival organization principles with the power and scalability of distributed databases.

Our position statement builds to these CAS investigations by suggesting a framework for "Levels of Computational Service" to better describe the emerging ecosystem and identify gaps and opportunities.

2. Levels of Computational Service

Journalists, researchers, planners, and other user patrons support their investigations with new methods of computational analysis. Libraries, archives, museums, and scientific data repositories hold data that will inform their disciplines. It is far easier today to analyze Twitter behavior than it is to investigate public life using public data from public institutions, such as government records, cultural heritage, and science data. We strive to make our public data and cultural memory as open to research as Twitter.

Computational analysis happens in various technical environments: on a single server; in distributed clusters; on cloud services. The tools we use have unique requirements, configurations, and hardware. It is said that a data stewardship organization cannot anticipate the uses for their data, but it is equally true that they cannot anticipate the tools used for analysis. Organizations need a service strategy that serves a range of users, from the most technically innovative, to the most time and resources constrained. We describe a range of services for collections as data without losing site of core services. This is a "maturity model" for stewardship organizations, with *levels of computational services* that show a clear progression toward full service.

2.1. Core Service Level

Shipping datasets into the researcher compute environment remains the critical use case, maximizing flexibility and allowing researchers to link many datasets into one corpus. Researchers need to *discover, scope, ship and make reference to datasets*. Though we may also move computational work across them, boundaries are an important place to define stable conditions, such as custody, provenance, security, and concise technical contracts. Even the most advanced repository must establish these boundary conditions.

- Define license terms, how can we use the data?
- Define provenance:
    - Who produced the data and why?
    - How did it arrive here?
    - Do versions exist elsewhere?
- Define dataset scope:
    - What makes the corpus complete?
    - Is it complete?

    - Is it growing? What is the update history?
- Transfer methods with integrity verification and resume from failure
- Persistently citable datasets

2.2. Protocols Service Level

- File-by-file transfer through HTTP API (instead of batch downloads, like ZIPs)
- Define citable subsets through custom queries or functions.
- Check for updates to any dataset or subset. (via HTTP API)
- HTTP API for navigation of structured collections:
    - Static site (Apache or Nginx auto-index of files)
    - Cloud Data Management Interface (CDMI)
    - Linked Data Platform (and Fedora API)
- Delivery to cloud and cloud-hosted, public datasets

2.3. Enhanced Service Level

- Derived data available as subsets:
    - plain text for documents and images
    - normalized file formats
    - tabular data for table-like sources
    - linked data for graph-like sources
- Machine-readable provenance records
- Crowd-sourcing of metadata
- Named entity indexing and subsetting (people, places, organizations, dates, events)
- Geospatial indexing and subsetting
- Consistent and citable random sample subsets (add random seeds to each observation)

2.4. Computer Room Service Level

Container technologies, such as Docker, ship a custom compute environment to the dataset location. A hosted database can be opened up for queries or distributed compute jobs. While not as flexible as the researcher environment, computer room services provide rapid and cost-effective analysis. Journalists on deadline benefit most from computer room services.
There are also growing calls, beyond the physical sciences, for analysis of big collections data in journalism and humanities scholarship. The sheer scale of big data makes transfer prohibitive, as is provisioning enough storage to host an entire corpus. At the Digital Curation Innovation Center at the University of Maryland's iSchool, we are actively developing the *DRAS-TIC* repository (Digital Repository at Scale that Invites Computation). Through *DRAS-TIC* we aim to deliver computer room-style services over heterogeneous digital collections and remove the limits of scale.

- Run an Apache Spark job on a defined dataset
- Host a compute container with a dataset mounted locally

- SPARQL query service
- Use techniques above to produce a new subset for transfer

3. Provisioning the Researcher Environment

From code notebooks to deployment scripts that provision clusters, it becomes easier to create and share compute environments. Research that aims towards publication will also need to track the research steps workflow. Through machine readable scripts and provenance, we can aim to reproduce an analysis at a different time and place, starting from the cited datasets and well described methods. The curation activities performed by a stewardship organization and the steps taken by the researcher can form an unbroken chain of events leading to a reproducible product.

Summary

For verifiable results in scholarship, or public trust in an independent press, we need to provide relevant datasets and services that make it straightforward to trace findings back to their source in the public record. We must confront a rightly skeptical reader, who faces increasingly high-flying visualizations and claims made from them. They are correct to demand links to the underlying evidence and methods. By providing these we enrich public understanding and trust. At the Digital Curation Innovation Center (DCIC) we have committed to this agenda and pursue it through our research projects, scholarly activities, and the active development of the DRAS-TIC software project, and the **building of a computational archival community** (see blog entry at: https://saaers.wordpress.com/2016/07/27/building-a-computational-archival-science-community/)