No Data, No Computation, No Replication or Re-use: the Utility of Data Management and Preservation Practices for Computational Journalism

Kris Kasianovitz
Stanford University Libraries
Green Library
Stanford, CA 94305
krisk@stanford.edu

Regina Lee Roberts
Stanford University Libraries
Green Library
Stanford, CA 94305
regirob@stanford.edu

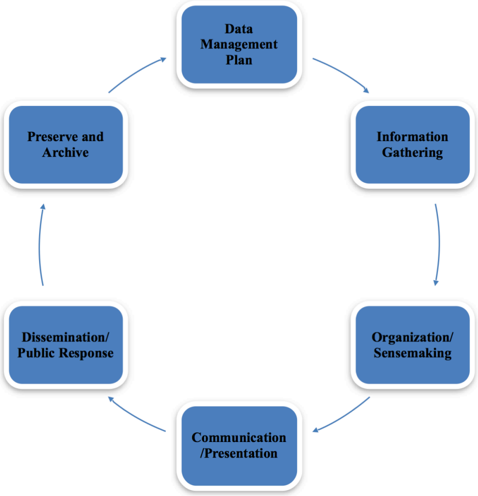
ABSTRACT

This paper tackles questions of data management, preservation and archiving that are critical to the *long view* of data access, use and re-use. This paper highlights important considerations for researchers and journalists within the context of ever changing methodologies for analyzing data for investigative reporting and computational journalism affecting public policy. As institutional repositories become more common and better established, there are more opportunities for corresponding data preservation and archiving of data in order to establish verifiable avenues for accountability and versioning in reporting. As computational journalism evolves, the care and access to data also needs to evolve. Libraries and librarians at research institutions are perfectly positioned to support access, preservation and management of data and some of the collateral information around analyzing data, such as models and algorithms. Included are examples of complex issues regarding data access, licensing and data archiving solutions with some recommendations for best practices.

Category: Data Management; Data Preservation; Data Archiving; Computational Journalism; Libraries
Theory: Methods
Keywords: Data Management, Data Preservation, Data Archiving, Libraries, Computational Journalism Methods

1. INTRODUCTION

The emergence of big data, open data, data science, computational journalism, and computational social science marks a shift in the way we produce, analyze and present the news, public policy and methods in social science research. In alignment with the deepening of methodological tool-sets to encompass this work, a shift in the way we think about access to data, and importantly, how we manage, preserve and archive the data is merited. The graphical representation that Diakopoulos[1] uses to describe the computational journalism processes is very similar to other data life-cycle graphs. However, it is missing these last three key components: management, preservation and archiving. We have added them in the figure below to create a more complete vision of how data can be managed in this domain.


Data Management Plan
Information Gathering
Organization/ Sensemaking
Communication /Presentation
Dissemination/ Public Response
Preserve and Archive

Dissemination and public response to news information necessitates that the data used to create reports are available in the long-term for re-use, replication and fact checking. As Cohen points out, “Data collection, management, and analysis have become ever more efficient, scalable, and sophisticated. The amount of data available to the public in a digital form has surged. Problems of increasing size and complexity are being tackled by computation [4]”. Without access to the same data or models that were used to write news stories and reports, where does that leave the readers, the public, and civil society? How can these stakeholder audiences be fully engaged in the dialogues, now and in the future? The answer is through preservation and archiving. This is why computational methods, be it in journalism or social science research, must also attend to thinking about democratization of the data, beyond the walls of data science, academia, and journalism.

With this shift in methodology, from the perspective of data librarianship, the trajectory of the data life-cycle for use in computational analysis is complex. Access to data for time-sensitive projects, such as news reporting or analysis of current events and issues should not be driven by what

[1] Presents a “process perspective” of journalism. [5]

special agreements and situations individual reporters/researchers have at hand or in place. The optimal situation would be to be able to select the data sets needed instead of having to modify projects because access to the *optimal* data is locked, restricted or prohibitively expensive. This is why preservation and archiving are critical components to be considered, even if it means more time and effort in order to do so. More repositories that include data, especially if it can be shared, also equates to better preservation.

2. THE LIBRARY PERSPECTIVE

The work of librarians is centered around providing broad, long-term (100+ years) access to information, in all formats. Librarians are interested in the democratization of information as a way of promoting civil society and equal access to many types of data sets, (which will be addressed later in this paper). Libraries along with institutional or academic departments are building institutional repositories to host data archives. Many institutions are now poised to actually archive and preserve local data sets. Ten or fifteen years ago, it was more of a concept, but now it is a reality taking shape. Because of this, services and support for data stewardship and curation are an expanding part of librarianship. In this role, librarians, data archivist, and bitcurators[2] work on acquiring and managing data used for large computational research projects. Ideally, faculty and students who either create or assemble data for their research and publications will also work with librarians to ensure that the data meets requirements of data management plans, federal and state policies requiring data deposit, and archiving of funded research. This includes working to understand issues of data acquisition; selection; access; curation; confidentiality; and preservation. Archiving of data is a positive step towards preservation, but it is not preservation unless it is backed up and has provisions for file versioning, migration and future imaging. The literature discussing data used in both computational journalism and investigative journalism is in alignment with what other academic researchers are seeking, collecting, and using. The data are very similar if not the same data from the same sources, (e.g. attorneys generals' opinions, legislation from fifty states, campaign finance, sea-level rise, prime and sub-prime mortgages, etc.).This observation, as well as actively working with these various data sets, gives us confidence that principles and practices used in academic library realm can and should be considered in the computational journalism realm [2].

3. WHAT ARE THE DATA?

Data are both structured and unstructured; texts, tables, .csv files, and unstructured; videos, photos, audio. More and more data are being created and captured at exponential rates; and there are more tools and opportunities to analyze larger corpora of data such as *big data* and *#opendata*. Thanks to the White House, (Executive Order 13642, 3 C.F.R. 2013 [1]), all the way down to city halls, open data policies have created wider access to government administrative data. So, it is easy to understand why Cohen et.al. writes about how large amounts of data plus computational power lets those who have access to the data and the knowledge of the computational tools to rapidly analyze and report on the analysis.[3,4]. In this era of rapidly expanding modes of research, access to open data is very critical to working on social issues and public policy initiatives and problems.

Even if an agency has not posted all of their data in one of these commonly used portals such as Data.gov, the open data and transparency policies of today greatly aid journalists, citizens, and researchers when requesting and getting access to this data, like the project that won the *Phil Meyer Award* that used 100 years of NOAA (National Oceanic and Atmospheric Administration) data [8]. Likewise, FOIA (Freedom of Information Act) and public records acts make obtaining government records possible[3]. Interviews and information collected in the field are also data sets with possible use cases for journalists. Data sets might be a "handful of individual items towards a news story" or "gigabytes of data" [4]. They can be in databases, flat files, or even handwritten documents.

Additionally, the realm of "big data" includes government data, but also scientific, commercial and private sources, e.g. sensor network data, social media, market research, *internet of things*, email, company records, large corpora of texts from books, journals, journal databases and even news sources. With web scraping tools and the *Internet Archive's Wayback Machine*, we have access to a diverse universe of "volume, velocity, and variety as vast as the Internet itself."[3]. Data documentation (metadata), scripts, api's, computational methods, algorithms, even the STATA do files are also considered part of the data which should also be made available for researcher, through systems of preservation and archiving.

4. GETTING THE DATA

Data can be completely "open" as in public domain or open access, see *Open Access Directory* and *GitHub*; or they can be restricted and even completely confidential or proprietary. All of these data characteristics or attributes impact access to or data purchasing, use, re-use, sharing and redistribution [2].

Obtaining data from any of the various governmental and non-governmental portals can be as simple as downloading a .csv file or using an .api. or via materials in the cloud. Granted, the latter of those does require some programming skill. More often, obtaining data requires a good deal of work and negotiation between two or more parties, be it a personal contact/source, a public records request, a phone call to an agency, or purchasing data. As Cohen et.al. points out, even though journalists are not bound by human subjects testing standards or research standards of peer-reviewed journals [3], librarians would advocate that researchers consider, whenever possible, certain aspects of

[2] Discusses bitcuration and digital forensics [6].

[3] It should be noted that obtaining access this way can be difficult and time-consuming.

these rules and standards when obtaining data, especially since this impacts redistribution and re-use of the data.

In libraries, we feel that if a researcher has gone through the difficulty to obtain and make data usable (paying for a large data dump from *WestLaw* or transcribing the city council public hearings) then the researcher should absolutely have a plan for managing, archiving, preserving with the intent to share this data. Researchers can help further research and save costs for future journalists, researchers, and the public by providing long-term access to their well-documented data. Since we are advocating for good data management practices in computational journalism, we encourage that the following be considered.

When obtaining data, it is beneficial to negotiate an agreement in the broadest possible terms for re-use and re-distribution if possible. This entails having explicit permission to place the data into a repository for others to use. Federal government information typically falls into the public domain so there are often few hurdles here. State and local government information, however, is actually copyrighted. Some statement of permission or even getting the agency to use a *Creative Commons* license declaration will clarify how the data can be used and distributed. Even though public record data is considered by most to be free and public once requested, having a clear statement about re-distribution is to everyone's advantage, especially if the data will go into an academic institutional repository. Obtaining these types of permissions is a common practice for academic librarians.

When data is proprietary or linked to subscription databases, journalists and researchers may find themselves needing to work with their institutional librarians in order re-negotiate license agreements that previously had not considered text and data mining methodologies. Academic researchers who negotiate license agreements that allows others to access the same data for different analysis or even replication are making ethical choices in relationship to access of their sources. Sometimes the only access allowed is the UI of a database when it has been opened up for metadata scraping. Additional data resources include: *Propublica*, *Factchecked.org, Public Insight*, and *NICAR*. Discipline specific repositories like LTER, Sloan, CUAHSHI and NICAR are making the data archiving process easier for deposit and access.

If data are prohibited from being redistributed, then document the sources, variables, and algorithms for analysis and archive that. This is important for two major reasons. First, making a large data set available without any information about the variables or how it was processed severely limits anyone's ability to replicate the work or re-use of the data. Secondly, in the cases of confidential or restricted data, researchers may not be able to share the actual data, but a description of the data, how it was obtained, and processed provides a valuable road map for anyone wanting to access similar data. [3].

Public affairs and accountability reporting can be challenging when local government agencies may or not be preserving and archiving their data. They are typically providing access to current data. The assumption is that agencies will be the stewards of their data; but we should be questioning that assumption as many state and local government agencies do not have long-term access, preservation and archiving on their radar; tape backup is their method.

Journalists especially those doing investigative work (watchdog work) are uniquely positioned to educate government agencies in the issue of access and archiving of their data. At this point in time, long-term access is not on most agencies' radars but it should be especially as open data portals and systems are developed. A great deal of the work done by government information librarians involves raising this awareness and trying to find solutions. If agencies hear this message from journalists, this can only help drive home the necessity of data archiving.

5. MANAGING AND SHARING DATA

In academia, grant funding agencies require researchers to have a Data Management Plan (DMP) which includes a strategy for depositing data. At Stanford University, we suggest using a tool such as the DMPtool[4]. While journalists are not required to have such a plan or even deposit their data, the tool can be a helpful guide for implementing some of the best practices suggested in this paper.

When it comes time to deposit the data that has been used there are several options. There area large archives such as the (Inter-university Consortium for Political and Social Research) ICPSR[5], which is a consortium repository for social science data and allows members to deposit and institutional repositories at colleges, universities and corporate institutions. Even news data centers or aggregated news providers have various degrees of depositor access and services. Organizations such as the Data Documentation Initiative (known as the DDI Alliance[6]) has been leading the effort to create an international standard for describing data from the social, behavioral, and economic sciences. This is an important step for the deposit of data. Data documentation of the methods, variables, analysis, R-SPSS-SAS codes, algorithms; method of acquisition are extremely important for future disseminated. This metadata is the data roadmap essential for re-use.

In a culture where using the most up to date and refreshed data are preferred, it would seem like preserving and archiving data are mundane tasks. Yet we argue that archiving for long-term access and preserving the data for future re-use is of critical value in the face of the common problems like "link rot", or server systems that only host the most recent version of a data set. What happens when someone wants to do a story about change over time and

[4] DMPTool; https://dmptool.org/

[5] ICPSR; https://www.icpsr.umich.edu/icpsrweb/landing.jsp

[6] DDI Alliance; http://www.ddialliance.org/

the only data available are current or the link in a paper or on a blog is defunct? One solution to the problem of link rot is being tackled by web archiving projects like Perma.cc[7]. This is a consortium archive allowing for archiving of web sites, articles, and other online objects. Researchers are allowed to submit the links to the pages that they wish to be archived after the links are reviewed they are preserved in a dark archive and then made available when people click on the perma.cc provided url. This is one effort to address the problems of link rot. An example of the importance of this is captured here in this statement:

"Suppose you're a law professor working on an article and you want to cite to the FBI's *Top Ten Most Wanted List*. Because that list changes over time, your citation is going to rot and your future readers won't be able to access your actual source at the original URL. That's bad for you and your readers [10]."

Likewise, backing up and storage is not preservation. Storing multiple copies of the data is always a good first step; so is using cloud storage for very large data sets. But this is just storage; storage of data doesn't necessarily get the same level of preservation. In academia, many researchers using services like Google Drive, Amazon Cloud, and GitHub[8].

But these services are not archive repositories. Ultimately, it is up to the individual to do this locally. This where an institutional repository or a consortium repository can play a crucial role[9] because of the commitment to such preservation strategies as "bit" checking and file migration.

6. CASE STUDIES

To emphasize some of the points in this paper, the following case studies are included.

6.1 Working To Develop Public Domain And Archiving Policies With San Mateo County Open Data Portal

Cities and counties are embracing open data. In California, most local governments are implementing open data portals powered by *Socrata*[10]. They have created data officer positions and in some cases, an entire department to get each of the agencies to add their data to the portal. San Mateo is engaging in this practice, quite wholeheartedly. They even hired an open data outreach liaison to work with agencies and the public to develop their portal: https://data.smcgov.org

Two key considerations from a data life-cycle and library perspective that are consistently missing in these portal implementations are: 1. Clarifying copyright and public domain for clear re-use and redistribution; and 2. Long-term access and archiving. At Stanford University Libraries, we've been fortunate to have contacts at the San Mateo County Information Technology department who have been listening to these issues. Through various informal meetings and email discussions, the county department is making sure to add a public domain declaration to their metadata. Additionally, we are embarking on a pilot program to try archiving the data sets into the Stanford Digital Repository. We are also discussing ways of managing archiving and preservation with *Socrata*. Because Stanford researchers are interested in the various county level data sets, this relationship is a step in the right direction. It also supports the work of the county in their civic responsibility of transparent government for the public. This type of partnering with libraries and consortium archives allows for greater open access to data, with greater control of long-term access. Just imagine how this might support the work of investigative reporters.

6.2 Proprietary Data In Subscription Databases For Text Mining:

Archiving and serving data is expensive, even when it is open data. There are real costs involved in the infrastructure and management of data. This is especially true when data is restricted by copyright or other concerns, it can be very difficult to obtain and impossible to share. Yet, sometimes the only source of data or the most *appropriate* data for a project is proprietary data.

With the increasing use of text mining tools and methods, researchers are looking to search corpora of text in order to analyze trends and, for example, conduct critical discourse analysis [7]. Gaining access to these texts is complicated by copyright, publisher's rights, and database aggregator license agreements, which almost always need to be updated to include data and text mining provisions. In these instances, web scraping without permission often violates copyright and institutional license agreements. At Stanford University Libraries, we have been working to negotiate an agreement for use of a corpus from a proprietary aggregated news database. The dialog has included discussions about creating a *secure virtual enclave* that would allow researchers to conduct experiments on a predetermined set of data on a server hosted by the vendor. This would allow the company hosting the aggregated data (which includes articles and transcripts) to stay true to their agreements with publishers, while allowing researchers an adequate amount of corpora to work with. This case has yet to be resolved and the current negotiations have been ongoing for almost a year. What this case emphasizes is the need for libraries or academic departments to negotiate license agreements or addendums to existing agreements that include options for text mining. This is especially true

[7] https://perma.cc/ [10]

[8] GitHub is a temporary space or parking lot for the data. See https://help.github.com/articles/can-i-archive-a-repository

[9] "Storage and Backup | Stanford University Libraries" Retrieved 12 August 2015, https://library.stanford.edu/research/data-management-services/storage-and-backup.]

[10] Socrata; http://www.socrata.com/

when negotiating new subscriptions to large aggregated journal article databases.

Another option would be for institutions to create a *secure virtual enclave* of their own that vendors could trust and agree to a data transfer for special projects. ICPSR has developed a working model of this idea for restricted social science data[11]. Granted, this is new terrain in not without risks for all parties involved. Yet, this shift in the way researchers want to work with these aggregated database collections really does require new models of access that need to be explored, negotiated and created. At Stanford, it has been really helpful to work with researchers in scoping out the specification and laying a foundation for future negotiations with vendors of data and aggregated journal databases.

RECOMMENDATIONS

Computational journalism will benefit from thinking through the data life-cycle including access, preservation and archiving. This will allow for validation, fact checking, and re-use by other journalists, researchers and the public.

Here are a few suggested best practices:
Develop a data management plan. Work this into early stages of research. This helps to structure a data collection and makes preservation and archiving easier. It also may help with securing grant funding - NSF requires this.
Document methods used. Even if using restricted data, the algorithms and scripts created by researchers along with details about the data set used, if archived, can help others to replicate a project in the future. It is becoming more common for these types of archived data and metadata to be included in the tenure process portfolio. Academics can benefit from working with their data librarians or data archivists to deposit data into their local institutional repository.

Seek out *secure* data storage. Again this may be available at the institutional level or through a consortium. Partner with cities, counties, federal agencies, or other non-profits, academic institutions in order to manage, archive, and preserve data. In some cases academic research libraries can respond to and support researchers who are engaged in this work. Discipline specific data archives are good options. Use *creative commons* licensing to let others use your data and encourage data creators to employ them as well for their own work. Educate the current and future generations of journalists to use these data management methods by incorporating the full data life-cycle concepts into training, in order to develop a new culture of data management for the field. Whenever possible, seek freely available resources for data management from universities and organizations.

The bottom line is, being able to link and access the appropriate data in meaningful ways in order to tell a story. Keeping this in mind, think carefully about access rights, preservation, and archiving for future computational journalists and researchers.

REFERENCES

[1] Executive Order 13642, 3 C.F.R. 2013. Making Open and Machine Readable the New Default for Government Information. (May 2013). Retrieved from: http://www.gpo.gov/fdsys/pkg/CFR-2014-title3-vol1/pdf/CFR-2014-title3-vol1-eo13642.pdf

[2] Borgman, C.L. 2015. *Big data, little data, no data : scholarship in the networked world /*. The MIT Press.

[3] Cohen, S., Hamilton, J. T., and Turner, F. 2011. Computational journalism. Communications of the ACM, 54, 10, (October 2011) 66. DOI=10.1145/2001269.2001288.

[4] Cohen, S., Li, C., Yang, J. and Yu, C. 2011.Computational Journalism: A Call to Arms to Database Researchers. (January 2011). Retrieved from; db.cs.duke.edu/papers/cidr11-CohenLiEtAl-cjdb.pdf

[5] Diakopoulos, N. 2011. A Functional Roadmap for Innovation in Computational Journalism. Retrieved from; http://www.nickdiakopoulos.com/2011/04/22/a-functional-roadmap-for-innovation-in-computational-journalism/

[6] John, J. L. 2012. Digital Forensics and Preservation. Digital Preservation Coalition. DPC Technology Watch Report 12-03 (November 2012). Retrieved from; http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.384.6486&rep=rep1&type=pdf

[7] Potts, A., Bednarek, M. and Caple, H. 2015. How can computer-based methods help researchers to investigate news values in large datasets? A corpus linguistic study of the construction of newsworthiness in the reporting on Hurricane Katrina. Discourse & Communication (March 2015), DOI=10.1177/1750481314568548.

[8] Reuters. 2015. What is data and computational journalism? (March 2015). Retrieved from; http://insideagency.reuters.com/2015/03/data-computational-journalism/

[9] Thompson-Kolar, M. 2014. ICPSR's Virtual Data Enclave Prepared to Accept New Restricted-Use Data. Retrieved from; http://www.icpsr.umich.edu/icpsrweb/ICPSR/support/announcements/2014/05/icpsrs-virtual-data-enclave-prepared-to

[10] Zittrain, J. and Albert, K. 2013. Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations. *SSRN Electronic Journal*. (2013). Retrieved from; http://dx.doi.org/10.2139/ssrn.2329161

[11] ICPSR website. Data Enclaves [9].