Understanding Computational Web Archives Research Methods Using Research Objects

Emily Maemura, Christoph Becker
Digital Curation Institute
University of Toronto
Toronto, Canada
e.maemura@mail.utoronto.ca,
christoph.becker@utoronto.ca

Ian Milligan
Department of History
University of Waterloo
Waterloo, Canada
i2millig@uwaterloo.ca

***Abstract*—Use of computational methods for exploration and analysis of web archives sources is emerging in new disciplines such as digital humanities. This raises urgent questions about how such research projects process web archival material using computational methods to construct their findings.**

This paper aims to enable web archives scholars to document their practices systematically to improve the transparency of their methods. We adopt the Research Object framework to characterize three case studies that use computational methods to analyze web archives within digital history research. We then discuss how the framework can support the characterization of research methods and serve as a basis for discussions of methods and issues such as reuse and provenance.

The results suggest that the framework provides an effective conceptual perspective to describe and analyze the computational methods used in web archive research on a high level and make transparent the choices made in the process. The documentation of the research process contributes to a better understanding of the findings and their provenance, and the possible reuse of data, methods, and workflows.

***Keywords*-computational methods; web archives; research objects; computational archival science; digital curation**

I. Introduction

The web has become a key source for scholars studying social and cultural phenomena. Web archives facilitate this work by documenting and preserving snapshots of the web for future research. In areas such as history, scholars have unprecedented opportunities to leverage access to massive primary sources of born-digital artifacts to make sense of individual experiences [1].

These new forms of archives trouble traditional conceptions of historical archival research, which assumes that any material found in archives is historically significant, having been assessed through archival appraisal. The scale of web archives confounds the traditional appraisal process. Web archives are replete with information that may not be significant to the research questions being asked.

The scale of these collections frequently mandates the use of big data analysis techniques to address their research questions. The adoption of these computational methods is a major shift for fields in the humanities. Approaches that have traditionally focused on close readings of source materials are moving to the use of distant reading methods [2].

We highlight three critical factors to be considered for humanities researchers using web archives:

1) *interrogating sources.* – Understanding how these web archives collections were created is key to judging the adequacy, appropriateness and limitations of the source material.
2) *understanding new methods.* – The use of emerging computational methods is a key prerequisite to working with large scale data sets.
3) *transparency of the research process.* – Findings are dependent on the validity of the computational processes and the adequacy of data.

These factors contribute to the need for a stronger methodological framework for research with web archives to understand the research process in more detail. This can serve as a common vocabulary for discussions of trust in the findings, as well as reuse of data, tools, or analytical techniques.

Our work combines perspectives of a digital humanist engaged in the computational exploitation of web archives and scholars working in the intersection of systems design and digital curation. Our joint interest in research with web archives led to an effort to structure research processes in order to develop detailed descriptions, with an eye to developing a research model. Our motivation for this study is to ask: *how do research projects in digital humanities process web archival material using computational methods to construct their findings?* We follow the work of other researchers in the digital humanities who have called for stronger frameworks for computational methods and the needs of researchers using web archives [3], [4].

To develop the framework, we draw on the concept of the Research Object (RO) developed to address the aggregations of resources and processes used in conducting computational research [5]. We use the RO profile as a structure to characterize three case studies that use computational methods to analyze web archives within digital history research, and discuss how the framework can illuminate issues such as transparency, provenance, and reuse.

The next section will outline background work in computational research with web archives and in conceptual frameworks to model computational science methods, with particular attention to Research Objects. We then develop a conceptual framework for Web Research Objects that we use to describe three case studies. These descriptions show that structured documentation of the research process contributes to a better understanding of the findings and their provenance. They serve as a basis for discussing questions of transparency, provenance, and the origins of the archival material used for analysis. These discussions show the value of the Research Object framework and raise a number of questions for further research.

II. Background

Web Archives and Historical Research

Web archives attempt to capture and preserve web data for future use and study. Different approaches to web archiving range from micro-archiving that captures details of the experience of interacting with individual web sites or elements to macro-archiving initiatives that use crawlers and other automated tools to capture web data [6], [7]. Major web archive initiatives for research use include the Internet Archive, national libraries and archives (such as the Library of Congress and the U.K. National Archives), academic libraries, as well as community efforts like the Archive Team. In parallel to archiving web pages, social media archives collect data from platforms such as Twitter or Instagram. They often rely on APIs from these services to construct their holdings.

For researchers in the social sciences and humanities, web archives are increasingly recognized as an essential source for studying cultural and social phenomena of recent decades [8], [9]. Niels Brügger has been studying the evolution of national domains [10]. Anat Ben-David has studied the now-deleted .yu domain and found parallels in the structure of the web archive with the break-up of the former Yugoslavia [11]. Richard Rogers' Digital Methods initiative has studied the Dutch blogosphere using the Wayback Machine [12], [13]. Matthew Weber used web archives to trace connections between local newspapers, studying a transformative moment in the history of American journalism [14]. The field is continually growing.

Working with web archival material presents researchers with opportunities for developing new approaches and methods of analysis, often because existing methods do not translate to them. For example, exploring web archives does not require navigating a finding aid, and collections can be sorted and filtered in multiple ways, resulting in multiple arrangements [15]. Unlike structured datasets from censuses, surveys, or statistical information, web archives are not temporally bounded nor predictable. Research methods for working with large-scale web data must take into account technical affordances like indexing, and different types of metadata available. Research with web archives frequently involves iterative cycles of text-mining or data-mining to sort and filter data, then applying analytics techniques like topic modeling, sentiment analysis, or network analysis. Whereas historians have traditionally found themselves wishing they had more information about the past, the "abundance" of web archives threatens to dramatically reshape forms of historical research and knowledge [16].

The emerging field of web archive research is therefore witnessing the development of new research methods. This is complicated by the nature of historical research, which often remains implicit in research methods. Indeed, many historical research projects involve an iterative process where methods develop emergently and remain undocumented. The variation across studies adds to the difficulty of understanding, comparing or validating results.

As a new form of archives, web archives challenge definitions found in traditional archival theory. Traditional approaches deal with aggregations of records that reflect specific business or institutional activities, and serve as evidence of those activities. In contrast, web archives often collect documents from a wide range of sources, created for different purposes, and formed around themes or events. The principles that guide the scoping of web archives may be seen as closer to curated collections in how they are selected and aggregated [17]. Interdisciplinary perspectives on provenance address chain-of-custody of objects [18], however the archival perspective is critical for supporting historical research by addressing the additional layers of meaning that context provides.

Previous work has investigated the practices of web archiving and researchers using web archives [3]. Studies done under the auspices of the Big UK Domain Data for the Arts and Humanities (BUDDAH) have echoed this, noting in particular the lack of guidance for humanist researchers [19]. Beyond these studies that ask researchers to reflect on their own methods, approaches are not often clearly documented. Instead, information about a study can exist in many places – code in a GitHub repository or embedded in a blog post, data occasionally published in an institutional repository or tweeted during intermediate stages, or methods articulated at practitioner conferences. For example, the historical article with little technical information is the final 'result', with code, technical approaches, and earlier presentations elsewhere. As time advances, too, platforms or technologies can change without warning, so that results generated with one version may not line up with newer versions.

A shared conceptual framework of the web archives research process is essential to systematize practices, advance the field, and to welcome new entrants to this area. It can also provide a shared vocabulary and flexible backbone to document and justify heterogeneous methods in a common perspective. Such a framework would be structurally useful to describe any research that investigates social questions based on web archives.

Frameworks from Computational Research in the Sciences

The Research Object (RO) framework structures data and computational workflows in reusable aggregates to facilitate reuse and reproducibility [5]. Its development was motivated by the need to support more sophisticated sharing of data and computational artefacts in computational science fields such as bioinformatics, medicine, astronomy, and genomics.[1] ROs aggregate the digital resources used in experiments in a structure that facilitates automated services, going beyond publishing results in scientific papers or datasets supported by linked data ontologies. Based on the claim that "linked data is not enough", the project aims to provide "a mechanism to describe the aggregation of resources", including data, computational environments and configurations, services used in carrying out an experiment or computational investigation [5].

Two aspects of the RO framework can be distinguished. On a research infrastructure development level, the RO framework implements a set of ontologies and mechanisms that enable the automated aggregation and linkage of named resources used in scientific investigations [20]. The primary aim is to increase the level of abstraction and semantic annotations to improve reusability and reproducibility in computational research. Some profiles are available that range from simulation experiments to computational workflow experiments [20].

On a conceptual level, the framework attempts to develop a perspective on the elements of computational research methods and articulate how the various pieces relate to each other. To do so, it also provides domain-specific standards for the component parts and relationships between them. Bechhofer et al. [5] describe the seven broad elements of a scientific study that a research object will include:

- **Questions** around a research problem, with or without a formal hypothesis. Descriptions or abstracts.
- **Organizational context** Ethical and governance approvals, investigators, etc. Acknowledgements.
- **Study design** encoded in structured documents
- **Methods** and scientific workflows or scripts, services, software packages.
- **Data** from observations or measurements organised as input datasets.
- **Results** from analyses or in silico experiments. Observations, derived datasets, along with information about their derivation or captureprovenance, algorithms, analyses, instrument calibrations.
- **Answers**. Publications, papers, reports, slide-decks, DOIs, PUBMED ids etc.

The authors also characterize the different types of reuse that the RO framework can help support. For example, they distinguish repurposing constituent parts of an RO (such as data or methods) in new experiments from reproducing the experiment using the same inputs and methods in an attempt to validate results. They also note the concept of ensuring an experiment is sufficiently documented to be 'replayable – to go back and see what happened which places requirements on metadata recording the provenance of data and results, but does not necessarily require enactment services. [5]

Reproducible computational research is a driving concern in science, and replication is often the ultimate test of a study's validity [21]. Support of reuse and reproducibility is also the focus of many data curation activities like the FAIR (Findable, Accessible, Interoperable, Reusable) Principles[2] and other guidance [22]. In contrast, computational reproducibility of results is not a goal for the humanities. However, the central principle of transparency mandates that the steps in a research process are 'recoverable' by others, i.e. that the representation of the applied research methods and techniques is sufficiently transparent for an independent observer to make sense of the research and arrive at a conclusion on how much confidence is warranted in its findings. While the RO framework was developed for work in the natural sciences, its focus on aggregating, structuring, linking, documenting and preserving elements of the computational process makes it a useful perspective to describe the archives research process.

III. Web Archives Research Objects: Three Cases

In search for a framework that can structure the complex aggregation of computational processes, services, forms of data, contexts, and approaches used in the research process with web archives, we adopt the RO concepts and leverage more mature perspectives on computational methods from the sciences. We seek to balance the systematic approaches from computer sciences with humanistic issues of provenance and trust. Therefore, less focus is placed on reproducibility as a test of validity, and more focus is placed on other forms of reuse.

Our approach uses RO framework concepts to structure a discussion of three cases. The examples are studies completed by our historian co-author, though they also involve interdisciplinary work with computer scientists and librarians. For each case we describe, characterize, and compare the processes and infrastructures used, the sources of data, the artefacts generated, and the important contextual factors that influence the overall study design.

Our process of applying the RO concept began with the seven elements of the research included in the RO: questions, organizational context, study design, methods,

[1] www.researchobject.org

[2] https://www.force11.org/group/fairgroup/fairprinciples

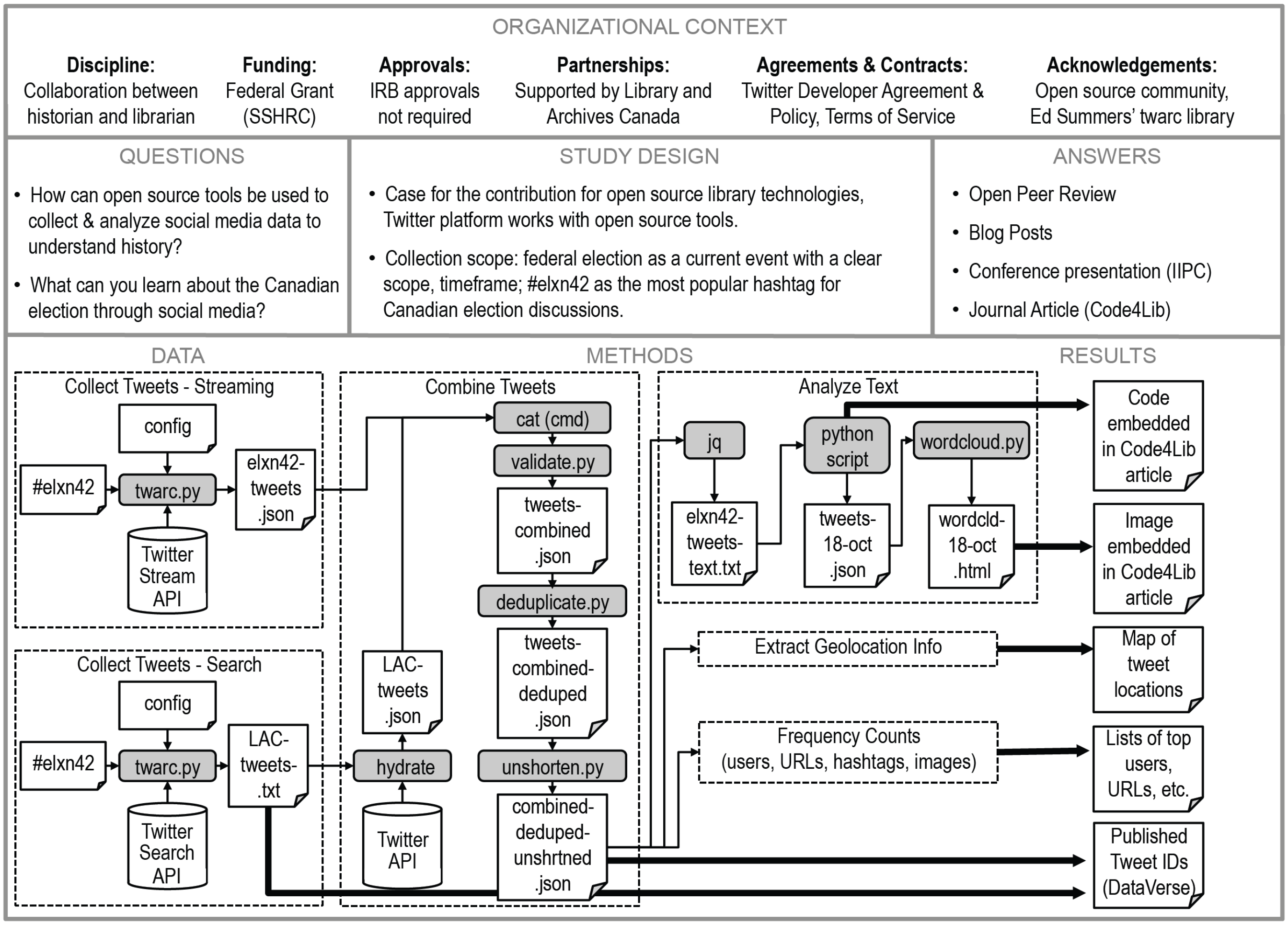
ORGANIZATIONAL CONTEXT
Discipline:
Collaboration between historian and librarian
Funding:
Federal Grant (SSHRC)
Approvals:
IRB approvals not required
Partnerships:
Supported by Library and Archives Canada
Agreements & Contracts:
Twitter Developer Agreement & Policy, Terms of Service
Acknowledgements:
Open source community, Ed Summers' twarc library
QUESTIONS
• How can open source tools be used to collect & analyze social media data to understand history?
• What can you learn about the Canadian election through social media?
STUDY DESIGN
• Case for the contribution for open source library technologies, Twitter platform works with open source tools.
• Collection scope: federal election as a current event with a clear scope, timeframe; #elxn42 as the most popular hashtag for Canadian election discussions.
ANSWERS
• Open Peer Review
• Blog Posts
• Conference presentation (IIPC)
• Journal Article (Code4Lib)
DATA
METHODS
RESULTS
Collect Tweets - Streaming
config
#elxn42
twarc.py
elxn42-tweets.json
Twitter Stream API
Collect Tweets - Search
config
#elxn42
twarc.py
LAC-tweets-.txt
Twitter Search API
Combine Tweets
cat (cmd)
validate.py
tweets-combined.json
deduplicate.py
tweets-combined-deduped.json
unshorten.py
combined-deduped-unshrtned.json
LAC-tweets.json
hydrate
Twitter API
Analyze Text
jq
elxn42-tweets-text.txt
python script
tweets-18-oct.json
wordcloud.py
wordcld-18-oct.html
Extract Geolocation Info
Frequency Counts (users, URLs, hashtags, images)
Code embedded in Code4Lib article
Image embedded in Code4Lib article
Map of tweet locations
Lists of top users, URLs, etc.
Published Tweet IDs (DataVerse)

Figure 1. Case 1: An Open-Source Strategy for Documenting Events (#elxn42)

data, results, and answers. We used these as a template to loosely structure a discussion of the process for each study, and ask questions about what happened including details not reported in publications. Answers were noted with an online whiteboard tool that provided a rough visualization drawing connections between elements. Further synthesis led to the structure used to describe outcomes for each case below.

A. Case 1: Open-Source Event Documentation

This case describes a project to create and analyze a web archive data set about the election of the 42nd Canadian Parliament in October 2015, focusing on the popular Twitter hashtag #elxn42 [23]. It is summarized in Fig. 1, which organizes the seven RO elements spatially.

The **organizational context** includes inter-disciplinary collaboration between a librarian (Nick Ruest at York University) and a historian (Ian Milligan at the University of Waterloo), both part of the Web Archives for Historical Research Group. The project was funded by the Canadian Social Sciences and Humanities Research Council (SSHRC). Data was shared with Library and Archives Canada (LAC). No ethics approvals were required since tweets are considered public data for research in a Canadian context. Notably, the collection and use of data is limited by the Twitter Terms of Service (Developer Agreement & Policy). Since the project focuses on open source tools, additional acknowledgments include the individuals who have worked to create the tools used, especially Ed Summers' twarc library [24].

The project aimed to enable libraries to collect social media data by demonstrating free and open-source techniques and methods available. Motivations for this study focused less on research **questions** or hypotheses, and instead it was an example of how social media can be used to understand historical events (both in general and the election).

The choices in the **study design** included scoping the project around the election, a recurring event with a clear time frame for data collection (starting with the election

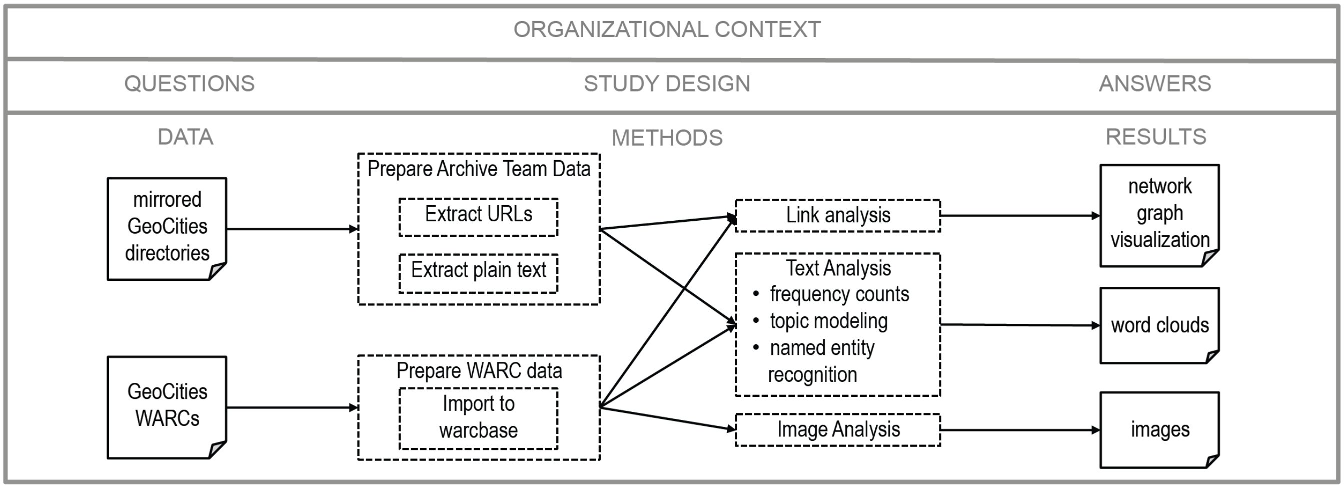

Figure 2. Case 2: GeoCities Community

announcement and ending with the swearing in of the new Prime Minister). Twitter was chosen due to its prominent role in political discussions, although its limitation of not being 'a representative sample of broader society' was noted [23]. *#Elxn42* was the most commonly used hashtag for the election. The study used open source methods and tools, and further embraced openness through an open peer review process and publishing in an open-access journal.

Data was both collected by the researchers, and a second dataset was received from LAC. Due to the restrictions of the Twitter Developer Agreement & Policy, extensive Tweet data cannot be shared between parties, and requires an additional step of 'rehydrating' with the Twitter API. The datasets were then combined using a series of tools from the twarc.py library. As noted in the papers analysis section, this means that tweets which were deleted after the time of collection would not be available for subsequent analysis.

Three different **methods** of analysis were applied to this combined dataset. First, tweets were grouped by day, and text analysis performed comparing word frequencies (visualized as a word cloud) using a python script. For tweets with geographic information, locations were interactively mapped. Quantitative analysis was performed on the data to identify top retweets, users, hashtags, links, most retweeted domains, and most frequent images.

As the study was concerned more with usefulness and feasibility of tools than testing or validating hypotheses, an important outcome is the publication of the code samples outlining the different processes represent the main outcomes of the study. Other **results** include wordcloud images, a map visualization, and top ten lists in the final article. The dataset of Tweet IDs is another result, published in DataVerse with a unique handle, but again requiring the Twitter API and 'rehydration' for future use and analysis.

The **answers** from this study were initially published as a draft online, available for a period of open peer review prior to publication in the openly accessible *Code4Lib Journal*. Additional conference presentations, blog posts, and the collected tweet IDs are publicly available.

B. Case 2: GeoCities Community

This project analyzed GeoCities.com, a platform hosting user-created sites that existed between 1994 and 2009 [25]. The GeoCities dataset is interesting as it represents an early example of mainstream website creation by a diverse group of users, built on a unique thematic neighbourhood structure. Fig. 2 organizes the seven RO elements spatially, but only the workflow is represented in detail here to save space.

For **organizational context,** the study was undertaken as a sole-authored research contribution by a historian, again funded by SSHRC. No ethics approval was required as all data was publicly available via the Wayback Machine.

The motivating **questions** asked here is if GeoCities users had a sense of community. A secondary question is what tools and approaches will historians need to study the web.

The **study design** included two iterations of analysis each using different datasets. The first, beginning in 2013, used the Archive Team torrent and a variety of tools. These approaches were further explored in a second round of analysis in early 2016 using a larger end-of-life dataset from the Internet Archive. This second round of analysis also took advantage of a web archiving analytics platform developed by the team (Warcbase). Both phases of analysis took an exploratory approach including analysis of links, text, and images, to understand what can be learned from methods of 'distant reading.' The study also included aspects of design research as the Warcbase tool was further developed via it.

Working with **data** provided by the Internet Archive required signing a research agreement limiting reuse and publication of the data. Additional data was provided by a publicly-accessible torrent of GeoCities, created by the web archivist group Archive Team. The two datasets together provided a foundation for the paper.

Two different sets of **workflows** were used in the different phases of analysis, in response to the differing types of data. Working with the Archive Team torrent (˜1GB) required additional steps of extraction and transformation of data to facilitate analysis. The torrent essentially reproduced all website directories to a local machine, from which URLs were extracted to create a network graph visualization with smaller sample sets of links. Similarly, plain text was extracted from the files for text analysis and topic modelling.

In the second round of analysis, these workflows were simplified since the data was structured in the standard WARC web archive container format. The data (˜4 TB) had to be loaded into HDFS on a cluster at the University of Maryland. Warcbase scripts, written in Scala and parsed via Apache Spark, performed extractive functions on the WARCs: extracting hyperlink graphs, full text, statistical URL analysis, and popular images via MD5 hashes.

The two input datasets also have varying documentation of how they were created. The Archive Team data was gathered with a wget command for geocities.com. Less information is available for the parameters of the Internet Archive crawl.

The analyses created a number of **results** artifacts and outputs: raw link structures, full text, popular images ordered by MD5 hash, counted URL statistics. Extracted results were then put into analytic platforms: topic modeling with MALLET, text analysis with bash and Mathematica scripts, image analysis and visualization with ImageMagick. Still, the output data itself doesnt directly answer the initial research question, as additional interpretation was required. For example, the visualization of network graphs reveals role of community leaders based on structure of links. The link analysis also reveals how webrings, guestbooks, etc. helped stitch interest groups together within GeoCities. Image borrowing reveals community people borrowed images from each other within discrete GeoCities communities. Topic modelling reveals that topics within GeoCities interest neighbourhoods lined up with what they should have—people made sites as we would expect them to.

The **Answers** from this study were published both informally in blog posts, and through conference presentations and in a forthcoming UCL volume. The research agreement signed with the Internet Archive limits the sharing or publication of specific derivative datasets.

C. Case 3: Content Selection and Curation

This case focused on the process of creating web archives, and choices of selection for what to include in the archive. Two web archives documenting the 2015 Canadian federal election are compared [17]. Fig. 3 represents the seven RO elements and focuses on depicting the workflow.

The **organizational context** for the project involves an interdisciplinary and inter-institutional collaboration as part of the Web Archives for Historical Research Group. Authors include the aforementioned librarian (Nick Ruest) and historian (Ian Milligan), joined by a computer scientist (Jimmy Lin at the University of Waterloo). The project was funded by SSHRC and the National Science Foundation (for Warcbase). No ethics approval was required. The project involved coordination with the University of Toronto Libraries for use of the Canadian Political Parties and Interest Groups (CPP) Archive-It collection (particularly Nich Worby, Government Information and Statistics Librarian). The paper relates to previous work involving these datasets, including the work described in Case 1, and previous work with the CPP collection.

The research **question** asked: What are the differences in coverage between different social media archiving strategies? The web archives compared represent two approaches to content selection and curation: an expert-selected list of sites versus a Twitter stream as a basis for URL selection. The **study design** set out to compare the content and coverage of the two collections. This included comparing which domains dominated each collection, as well as an intersection analysis of all URLs between collections. URLs from each collection were also checked for inclusion in the Internet Archives broader crawls for the time period.

Three different sets of **input data** were used, and each had varying levels of documentation describing how they were generated. Twitter data was collected by the authors and did not require further explanation. However, as noted in Case 1, Twitter terms of service constrain data sharing. The CPP collection was opaque as the set of 50 seed URLs for the collection was established in 2005 with little documentation, by a subsequently retired librarian. The seed list is periodically updated by the current librarian. The Internet Archive collection approach is similarly opaque, with little documentation. Previous research examined coverage [26].

The **methods** used included preparation of the Twitter data by unshortening and extracting all URLs, then deduplicating for a set of unique URLs. These Twitter-curated URLs were also crawled to create a separate web archival collection for future use. Quantitative analysis of the Twitter and CPP URLs identified the top domains represented in each. Then an intersection analysis was completed with a bash script, resulting in the percentage of Twitter URLs found in the CPP collection and vice versa. The intersection analysis for each collection with the Internet Archive involved querying the unique URLs for each collection with the Wayback CDX Server API, and similarly determining the percentage of coverage. Further, for the comparison of Twitter with the Wayback Machine, the top ten domains included and excluded from Waybacks collection were determined.

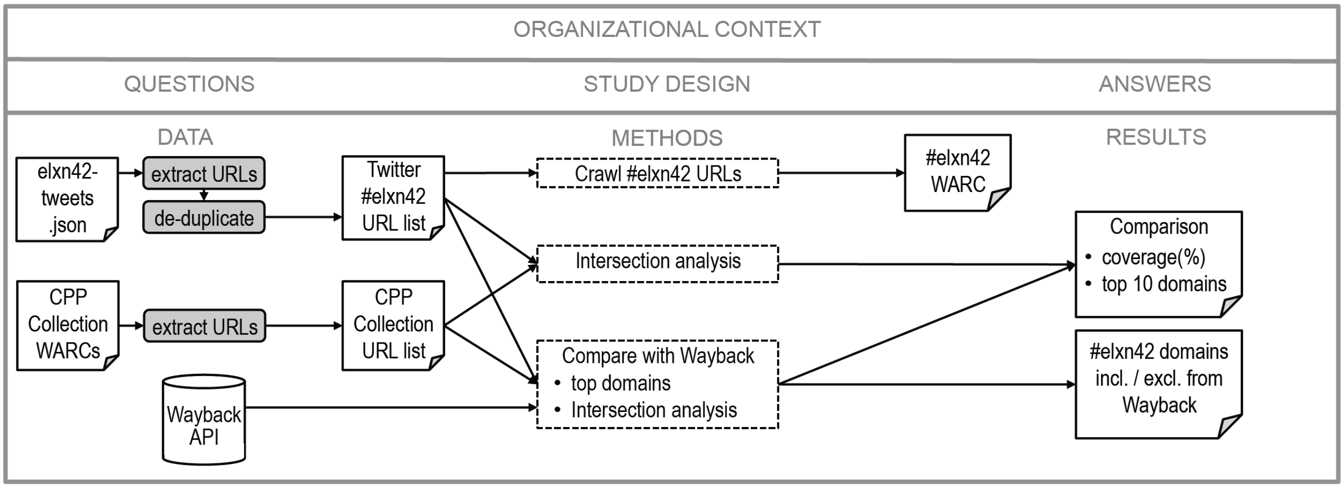

Figure 3. Case 3: Content Selection an Curation - The Gatekeepers vs. the Masses

The **results** of the intersection analysis provide a quantitative indication of the coverage between collections. The low level of intersection of Twitter URLs found in the CPP collection (0.269%) and CPP URLs found in the Twitter collection (0.341%) led to a recommended hybrid selection and curatorial approach. Comparing the Twitter collection to the Internet Archive found 10.06% coverage compared to a 74.30% coverage of CPP URLs in the Wayback Machine.

The **answers** were published as a conference paper and presentation. Preliminary work appeared in a blog post, which served as an impetus for further research [27].

IV. Discussion

Towards a RO Profile for Web Archives Research

One outcome from studying these cases is the development of a preliminary profile for a Web Archives Research Object. This framework takes the seven elements of the RO (questions, context, study design, methods, data, results, answers), and expands on the relevant or influential factors to identify for each. These are listed in Table I. Some aspects required adaptation to domains outside the natural sciences.

Organizational context was found to be of primary importance since it impacts the overall approach. Disciplinary perspectives are especially relevant to note since research with web archives can involve researchers from fields such as history, library and information science, computer science, media studies, communication. Work taking place in a collaborative team may involve a variety of roles, as well as other forms of support or partnerships. The processes required to ensure ethical compliance may also vary depending on disciplinary perspectives.

A notable finding is the importance of other legal frameworks, agreements and contracts that may influence the degree to which the sources, intermediate artefacts and results can be made available. For these cases this included the Twitter Terms of Service and Developer Agreement, as well as the Research Agreement with the Internet Archive.

It is also important to understand the decisions in selecting source data, if other options were considered but not useful or not available. Study design should also address the reasoning behind choices in research methods.

Generally, greater documentation is needed on **methods, data, and results**, as well as the relationships between them. If the study included collecting data, the use of specific services like APIs should note the versions and timeframes. If the study uses data generated previously, questions arise of where and when it was collected, and by whom. As Borgman notes, the greater distance between the scholar and the origins of the data means that interpretation relies on ‘formal knowledge representations’ [28]. This could include metadata or other descriptions of how data was generated.

Separate workflows are often required for preparing data prior to analysis. This includes integrating datasets from different sources, as well as other forms of selecting and filtering. The analysis workflows explored here separate along the lines of format: plain text, images, and platform-specific data structures (e.g. hashtags, in-links or out-links). The results of analysis take many forms, like derived datasets or other artifacts like visualizations. A notable tension between the original development of ROs for the sciences and an application to humanities research is how or where to acknowledge the critical step of *interpretation* of the artifacts generated by computational processes. We have addressed these as part of the **Answers** i.e. in the text of publications, papers, reports or other presentations of the research, and can also include alternate forms of knowledge mobilization.

The approach also supports a much-needed conversation about reuse of computational workflows and methods.

Table I
AN INITIAL PROFILE FOR WEB ARCHIVES RESEARCH OBJECTS

RO Element	Concerns for Web Archives RO profile
Organizational context	Note which *disciplinary perspectives* are represented, especially in interdisciplinary collaborations – these may be reflected in *Funding Agency* support. *Project team, investigators* should be identified and *other roles* acknowledged, including institutional or individual *partnerships*. In addition to ethical and governance approvals, also consider which other *legal Agreements and Contracts* impact data use and sharing, and any relevant jurisdictions that determine what web data is public.
Questions	Include *motivations* for the study such as the overall *research contribution*. For interdisciplinary work it is also useful to consider if questions are framed towards certain *audiences* - is the work relevant for researchers or practitioners in a certain field?
Study design	Since designs can vary widely by discipline, it may be useful to consider *conceptual perspectives* reflected in research questions (e.g. positivist or interpretivist approaches). Other relevant decisions include how data sources were selected, which types of questions were considered, how the scope of the study was determined and any limitations.
Methods	Identify specific scripts, services (like APIs), and software packages used. Was code deposited, and if so, in what repository?
Data	Include information about *sources* (single or multiple) and how data was collected directly by the research team or by others. What was the *timeframe* of collection, and was it contiguous? Which *formats* are available, and are they interoperable?
Results	Consider if results are Findable, Accessible, Interoperable, Reusable (FAIR). How are they published and made citable? How were they validated?
Answers	In addition to identifying formal and informal publications, consider the methods used for *interpretation of results*, which is an essential part of humanities scholarship.

Consider Table II, which examines the elements of Case 1 and the degree to which they are findable, accessible, interoperable, and reusable. The analysis highlights several aspects of this project that are problematic in terms of its reusability. While Case 1 is extensively documented in the published article, there remain elements of the study that were not documented like runtime environment, logs, and code repositories. Determining the level of detail necessary to document about workflows also requires consideration of what types of reuse are intended, and possible. While reusability is much more complex than this reductive list, and the structured checklist is far from the advanced state of art in reproducible computational science, it is efficient and quick to complete as a first step to identify opportunities for improvement. The researcher found it very helpful in understanding the nature of his own research.

How the RO Profile can support the Research Process

This framework can support the research process by helping scholars in emerging disciplines such as the digital humanities to document their practices systematically. Even within a single researcher's set of cases, there is a broad variety of sources that shift over time and tools that rapidly evolve. The consistent structure of describing these components helps in keeping track of these shifts and to make sense of how the findings are constructed. At the same time, however, this structure provides only a preliminary understanding on a high level, and needs to evolve through further iterations and extended studies. It must be complemented by a better understanding of the relationships between the components.

Applying ROs to the research process allows several significant benefits for the emerging field of web archiving. First, we can more efficiently develop and compare research projects. Instead of scholars individually developing research practices, a set of common practices, workflows and approaches can be developed and reused. The RO framework provides a view of the interrelated parts, and can encourage more systematic processes of publication and sharing of code, saving intermediate datasets, and preserving important organizational such as research agreements and Terms of Service. Second, in a new and developing field, the increased transparency of methods allows for both new entrants to the field as well as a shared pooling of technical knowledge in a largely non-technical community of scholars. Such transparency facilitates the sharing of datasets and workflows, allows the citation of technical elements, and the recognition of non-traditional venues such as code repositories, blog posts, and lightning talks.

In applying these concepts, we are keenly aware that the computational sciences are more oriented towards replication than the exploratory and interpretive work in the humanities. Still, the approaches to reuse of specific tools or workflows is relevant to working with web archives, and we believe that using common frameworks can help foster collaboration between computer sciences and the social sciences and humanities.

Beyond reuse and sharing that directly supports the research process, the framework can also be used to support peer reviewing, as a checklist or scaffolding to begin assessing aspects of this emerging field. This framework could also be used to explore cases across disciplines, similar to Borgman's work [28] that facilitated better understanding of the role research data play in digital scholarship across different disciplines. These discussions are needed to begin sketching out a methodology for web research, separating the often-conflated aspects of tools, methods, and methodologies, as well as identifying the diverse (and sometimes incompatible) guiding principles, types of research, and approaches.

V. CONCLUSION

Computational methods are emerging for exploration and analysis of web archives sources, but these are not clearly

Table II
HOW FINDABLE, ACCESSIBLE, INTEROPERABLE, REUSABLE ARE THE COMPUTATIONAL METHODS IN CASE 1?

Element	**F**	**A**	**I**	**R**
Workflow: Collect Tweets				
Input: **#elxn42 hashtag** This hashtag was selected as the main object of research.	Y	Y	N	Y
Script: **twarc.py** This open source Python module was developed on Github. The used version is v0.7.0 at https://github.com/edsu/twarc/releases/tag/v0.7.0.	Y	Y	Y	Y
Parameters: **twarc Configuration Settings** Search parameters are set with a series of key-value pairs. The exact settings used for collection are not documented or publicly available. If saved they would be interoperable for other collections using the twarc library.	N	Y	Y	N
Service: **Twitter Search API** The Twitter API allows users to access datasets from Twitter users, with specific limits - the Twitter policy does not allow more than 1% of the publicly accessible datastream. The Search mode runs a search for all tweets with a given hashtag, going back roughly 5-7 days. LAC ran it 1-2 times/week. The version used is no longer operational.	Y	N	Y	N
Result: **Tweet datasets - JSON** The tweets collected directly by the researchers were saved as a JSON dataset. The Twitter policy does not allow sharing of this data or making it public in this form. *Reuse is possible by the researcher who collected the data.	N	N	Y	N*
Result: **Tweet datasets - IDs .txt** The tweets collected by Library and Archives Canada were made available as a dataset of only the Tweet IDs in text format. They are hosted on ScholarsPortal DataVerse, with documentation, and can be used to 'rehydrate' a Twitter dataset. [https://dataverse.scholarsportal.info/dataset.xhtml?persistentId=hdl:10864/11310]	Y	Y	Y	Y
Workflow: Combine Tweets				
Script: **Twarc.py –hydrate** Python module command takes a list of tweet IDs and 'rehydrates' them by returning the full JSON of each tweet. The version used was Twarc.py v.0.7.0 . This depends upon the ***Service:*** **Twitter API**	Y	Y	Y	Y
Command: **cat** The cat command was used to concatenate (combine) the tweets gathered by the LAC and York/Waterloo teams.	n/a	n/a	Y	Y
Script: **validate.py** Python command ensures that each line of the .json files is a valid JSON object. (part of twarc.py v.0.7.0 library)	Y	Y	Y	Y
Script: **deduplicate.py** Python command deduplicated the tweets, so that only unique Tweet IDs remained. (part of twarc.py v.0.7.0 library)	Y	Y	Y	Y
Script: **unshorten.py** Python command found all shortened URLs and expanded them for the dataset. (part of twarc.py v.0.7.0 library)	Y	Y	Y	Y
Result: **JSON files** A series of JSON files were created at various stages of this process. Due to Twitter Terms of Service, this file cannot be shared as it contains the full JSON. *Reuse is possible by the researcher who collected the data.	N	N	N	N*
Result: **Tweet datasets - IDs .txt** The final set of combined tweets were made available as a dataset of only the Tweet IDs in text format. They are hosted on ScholarsPortal DataVerse, with documentation, and can be used to 'rehydrate' a Twitter dataset. [https://dataverse.scholarsportal.info/dataset.xhtml?persistentId=hdl:10864/11311]	Y	Y	Y	Y
Workflow: Analyze Text				
Script: **jq** The command-line JSON 'jq', was used to extract the plain text of every tweet. JQ is open source available on GitHub	Y	Y	Y	Y
Script: **custom Python script** The researchers wrote a Python script to sort each tweet by day and created separate text files for each day. The script was provided in the text of the open-access Code4Lib article.	Y	Y	Y	Y
Script: **Wordcloud.py** Using wordcloud.py, we created a word cloud of tweet text for each day. This allowed us to track major shifts in tweet content. (part of twarc.py v.0.7.0 library)	Y	Y	Y	Y

or comprehensively reported in publications. This makes it difficult to compare, assess, and validate results. Historians have rarely made their methods clear and instead rely on the tacit implicit knowledge of traditional approaches. The increased use of computational workflows highlights the need to enable such scholars to document their practices systematically to improve the transparency of their methods.

The past decade has seen great advances on such crucial components as reproducibility in computational science, the study of computational research methods, techniques and models that facilitate the reuse of computational workflows, and mechanisms to facilitate the sharing data and workflows.

By uniting these perspectives, this article aimed to contribute to a shared understanding of emerging methods through applying the Research Object concept in three cases of computational web archival research. The initial results provide a stepping stone to a common framework for characterizing such research. Bundling constituent parts together enabled researchers to make sense of decisions, approaches, and findings.

Our current work builds on this in several ways:

1) Expand the set of cases to describe a wider variety of research scenarios and derive an effective template to support researchers in documenting their methods.
2) Expand the RO model to include more information on the origin of the data itself. Rather than taking data as an input for granted, with web archives the RO needs to include details on the scope of the web crawl, the technical decisions made, websites excluded, and beyond.
3) Explore, through the lens of Web Research Objects, how the design of cyber-infrastructures for archive-based computational research such as web crawlers and social media archives can support more transparent research methods and outcomes.

ACKNOWLEDGMENT

Part of this work was supported by the National Science and Engineering Research Council (NSERC) through RGPIN-2016-06640, and the Social Sciences and Humanities Research Council (SSHRC) through Insight Grant 435-2015-0011 and Canada Graduate Scholarship 767-2015-2217. Ian Milligan was also supported by the Marshall McLuhan Centenary Fellowship in Digital Sustainability at the University of Toronto iSchool Digital Curation Institute.

REFERENCES

[1] I. Milligan, "Lost in the Infinite Archive: The Promise and Pitfalls of Web Archives," *Intl J Humanities &*

Arts Computing, vol. 10, no. 1, pp. 78–94, Mar. 2016. [Online]. Available: http://www.euppublishing.com/doi/abs/10.3366/ijhac.2016.0161

[2] F. Moretti, *Graphs, Maps, Trees: Abstract Models for Literary History*. Verso, Sep. 2007.

[3] M. Dougherty and E. T. Meyer, “Community, tools, and practices in web archiving: The state-of-the-art in relation to social science and humanities research needs,” *JASIST*, vol. 65, no. 11, pp. 2195–2209, Nov. 2014. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/asi.23099/abstract

[4] C. L. Borgman, “The Digital Future is Now: A Call to Action for the Humanities,” vol. 3, no. 4, 2009. [Online]. Available: http://www.digitalhumanities.org/dhq/vol/3/4/000077/000077.html

[5] S. Bechhofer, I. Buchan, D. De Roure, P. Missier, J. Ainsworth, J. Bhagat, P. Couch, D. Cruickshank, M. Delderfield, I. Dunlop, M. Gamble, D. Michaelides, S. Owen, D. Newman, S. Sufi, and C. Goble, “Why linked data is not enough for scientists,” *Future Generation Computer Systems*, vol. 29, no. 2, pp. 599–611, Feb. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X11001439

[6] J. Masanès, *Web Archiving*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. [Online]. Available: http://link.springer.com/10.1007/978-3-540-46332-0

[7] N. Brügger, *Archiving Websites: General Considerations and Strategies*. Aarhus: Center for Internetforskning, 2005.

[8] S. M. Schneider and K. A. Foot, “The Web as an Object of Study,” *New Media & Society*, vol. 6, no. 1, pp. 114–122, 2004.

[9] N. Brügger and R. Schroeder, Eds., *The Web as History*. London: UCL Press, 2017. [Online]. Available: https://www.ucl.ac.uk/ucl-press/browse-books/the-web-as-history

[10] N. Brügger, D. Laursen, and J. Nielsen, “Studying a nation’s web domain over time: analytical and methodological considerations,” Palo Alto, California, Apr. 2015. [Online]. Available: http://netpreserve.org/sites/default/.../2015_IIPC-GA_Slides_02_Brugger.pptx

[11] A. Ben-David, “What does the Web remember of its deleted past? An archival reconstruction of the former Yugoslav top-level domain,” *New Media Society*, p. 1461444816643790, Apr. 2016. [Online]. Available: http://nms.sagepub.com/content/early/2016/04/27/1461444816643790

[12] R. Rogers, *Digital Methods*. Cambridge, Mass: MIT Press, 2013.

[13] E. Weltevrede and A. Helmond, “Where do bloggers blog? Platform transitions within the historical Dutch blogosphere,” *First Monday*, vol. 17, no. 2, Feb. 2012. [Online]. Available: http://firstmonday.org/ojs/index.php/fm/article/view/3775

[14] M. S. Weber, “From Big Data to Big Theory: Lessons Learned from Archival Internet R,” Atlanta, Georgia, 2016. [Online]. Available: http://www.slideshare.net/mwe400/from-big-data-to-big-theory-lessons-learned-from-archival-internet-research

[15] J. Bailey, “Disrespect des Fonds: Rethinking Arrangement and Description in Born-Digital Archives,” *Archive Journal*, no. 3, 2013. [Online]. Available: http://www.archivejournal.net/issue/3/archives-remixed/disrespect-des-fonds-rethinking-arrangement-and-description-in-born-digital-archives/

[16] R. Rosenzweig, “Scarcity or Abundance? Preserving the Past in a Digital Era,” *The American Historical Review*, vol. 108, no. 3, pp. 735–762, Jun. 2003, articleType: research-article / Full publication date: June 2003 / Copyright 2003 American Historical Association. [Online]. Available: http://www.jstor.org/stable/10.1086/529596

[17] I. Milligan, N. Ruest, and J. Lin, “Content Selection and Curation for Web Archiving: The Gatekeepers vs. the Masses,” in *Processings of the Joint Conference on Digital Libraries*. Newark, New Jersey: ACM, 2016.

[18] V. L. Lemieux, Ed., *Building Trust in Information*, ser. Springer Proceedings in Business and Economics. Cham: Springer International Publishing, 2016. [Online]. Available: http://link.springer.com/10.1007/978-3-319-40226-0

[19] R. Deswarte, “Revealing British Euroscepticism in the UK Web Domain and Archive Case Study,” Jul. 2015. [Online]. Available: http://sas-space.sas.ac.uk/6103/#undefined

[20] K. Belhajjame, J. Zhao, D. Garijo, M. Gamble, K. Hettne, R. Palma, E. Mina, O. Corcho, J. M. Gmez-Prez, S. Bechhofer, G. Klyne, and C. Goble, “Using a suite of ontologies for preserving workflow-centric research objects,” *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 32, pp. 16–42, May 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1570826815000049

[21] G. K. Sandve, A. Nekrutenko, J. Taylor, and E. Hovig, “Ten Simple Rules for Reproducible Computational Research,” *PLOS Comput Biol*, vol. 9, no. 10, p. e1003285, Oct. 2013. [Online]. Available: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003285

[22] A. Goodman, A. Pepe, A. W. Blocker, C. L. Borgman, K. Cranmer, M. Crosas, R. D. Stefano, Y. Gil, P. Groth, M. Hedstrom, D. W. Hogg, V. Kashyap, A. Mahabal, A. Siemiginowska, and A. Slavkovic, “Ten Simple Rules for the Care and Feeding of Scientific Data,” *PLOS Comput Biol*, vol. 10, no. 4, p. e1003542, Apr. 2014. [Online]. Available: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003542

[23] N. Ruest and I. Milligan, “An Open-Source Strategy for Documenting Events: The Case Study of the 42nd Canadian Federal Election on Twitter,” *The Code4Lib Journal*, no. 32, Apr. 2016. [Online]. Available: http://journal.code4lib.org/articles/11358

[24] E. Summers, “edsu/twarc,” 2016. [Online]. Available: https://github.com/edsu/twarc

[25] I. Milligan, “Welcome to the Web: The Online Community of GeoCities and the Early Years of the World Wide Web,” in *The Web as History*, N. Brügger and R. Schroeder, Eds. London: UCL Press, 2017.

[26] S. G. Ainsworth, A. Alsum, H. SalahEldeen, M. C. Weigle, and M. L. Nelson, “How Much of the Web is Archived?” in *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, ser. JCDL ’11. New York, NY, USA: ACM, 2011, pp. 133–136. [Online]. Available: http://doi.acm.org/10.1145/1998076.1998100

[27] I. Milligan, “Institutional vs. Twitter Seed Lists for Web Archives,” Nov. 2015. [Online]. Available: https://ianmilligan.ca/2015/11/26/institutional-vs-twitter-seed-lists-for-web-archives/

[28] C. L. Borgman, *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, Mass.: MIT Press, 2015.