Archiving Database Driven Websites for Future Digital Archaeologists: The Archiving of TAPoR

CSDH 2020

Bennett Kuwan Tchoh, Geoffrey Rockwell

Introduction

With the increase of information technology access and use, digital born internet projects are very common these days. Many projects are built that provide a wide variety of services—data collection, data analysis, communication, collaboration, etc. Most of these projects manipulate data which are stored in databases and the use of the services of the website itself generates new data which are stored in a database. In the digital humanities (DH) field and in other fields of study, projects are constantly being started, some are completed while others end up being abandoned but of these projects, whether they were completed or not, very few of them are archived (Brown et al., 2009). If lucky, sometimes we can see traces of discontinued internet based projects from internet archiving websites. These websites were developed to prevent the loss of records of our activities in the World Wide Web, but the amount of information available is usually limited. This lost information is part of our digital heritage meaning part of our online history is lost.

Archiving the web is a complex task with many challenges and it seems such projects can be run only by big institutions like government run institutions and big digital libraries like the Internet Archive which runs the Wayback Machine (a digital archive available at https://web.archive.org/). Archiving the internet in its entirety is an impossibility considering the

exponential growth of data and the fact that the deep web remains inaccessible (Kitchin, 2014; Masanès, 2005). These organisations will usually use web crawlers and bots to periodically scrape and save internet pages. Although the mass archiving of internet data produces a wealth of information, the automatic nature of web scraping results in much information being left out, especially dynamically generated content hidden in databases. On the other hand, sensitive information and copyrighted material are copied indiscriminately. Lecher (2006) proposes smaller scale archiving of a particular topic of interest as a solution to the problems faced by large scale archiving like the Internet Archive. His solution involved a university or department or a museum setting up both the hardware and software for an archive on a particular topic or topics and making the data available to open access.

The need for preservation of information has been largely recognised and it is common these days for institutions and universities to have their own repositories in which they store data produced from the research undertaken under the institution or university. In recent years, there has been an exponential growth in the number of institutions having their own repositories. These repositories sometimes have dedicated staff and sometimes the staff is part of the library section. They usually have hardware and IT personnel to maintain the data available online at all times. Most of these repositories work through a process of self-archiving (SA) in which the authors themselves deposit their research data in prepared collections. Most universities require that students submit digital copies of their thesis. The repository personnel or the students themselves deposit their thesis in the repository. Extracting metadata from the thesis is very straightforward and the theses are saved such that they can easily be found. This is a human run process and the data is saved efficiently in the right category or collection, unlike in automated processes with the use of crawlers that create much room for errors. In present day,

academic repositories are interlinked such that a single search query on an academic search engine will search a large number of repositories.

For most of the projects that are archived, the archiving process usually starts at the end of the project. This makes creating the archive more difficult considering that the archivist will have to review the entire project and select and prepare the information to be archived. Starting the archiving process at the end also means that there are high chances that at least some of the data on the previous stages of the project would have been lost

Better late than never, this paper will discuss the process of self-archiving TAPoR a database driven website which is presently in its third iteration.TAPoR stands for Text Analysis Portal for Research and was originally a multi-institutional CFI-funded project that brought together 6 universities across Canada to develop digital humanities infrastructure both at the universities and shared.[1] The portal itself was an example of shared infrastructure that was meant to provide a vertical portal of text analysis services to humanities researchers. (Rockwell 2006) The first version of the portal combined both tools for discovering tools with actual text analysis tools and other related social media services. The portal was developed by Open Sky Solutions under the leadership of James Chartrand and went from public beta to full release in 2007. In later versions the portal was redeveloped and focused on the discovery of tools so that the third version (version 3.0) doesn't provide access to tools, but does link to tools, does maintain historic information about tool projects, does provide code snippets and does encourage the exploration of a range of different types of tools of interest to humanists. The

[1] The universities included the University of Victoria, the University of Alberta, the University of Toronto, McMaster University, l'Université de Montréal, and the University of New Brunswick. The project was led by Geoffrey Rockwell who at that time was at McMaster and was funded in 2002.

actual text analysis tools were reimplemented in a separate project, the Voyant project (http://voyant-tools.org) led by Stéfan Sinclair (Rockwell & Sinclair, 2016).

In 2018 the data from the DiRT (Digital Research Tools) project from UC Berkeley was absorbed by TAPoR as that project had lost support (Grant et al., 2020). In 2019 TAPoR started a collaboration with the DARIAH SSH Open Marketplace sharing our data with that European project.[2] The fragility of projects that act as discovery services like DiRT, TAPoR and the Open Marketplace made it important that, at the very least, we archive our data at this crucial junction after we had absorbed the DiRT data. Even if the infrastructure is discontinued, the data should be capable of being accessed.

Generalities on archiving

Archiving can be defined as the process of moving files that are not frequently used from a primary memory location into a cheaper long-term retention location (Reeve, 2013). It has been common practise for universities to store the products of their research in departments or libraries. The first data archives which were established in the United States in the 1960s were different from what we consider data archives today and stored mainly survey data from the social sciences. Prior to that, computerised data (usually in the form of punch cards) were increasingly being stored and it was realised that these data were valuable for they were generally underutilized, they needed to be checked (considering that many conclusions could be drawn from them) and will have value for historians (Doorn & Tjalsma, 2007). In more recent times, it has been noticed that the infrastructure available for researchers to deposit their data for open access is not as developed as infrastructure available for them to deposit their publications from these data (Brody et al., 2007); added to that is the fact that many researchers

[2] See https://www.sshopencloud.eu/ssh-open-marketplace

are reluctant to make their data publicly available and some of the factors that are associated with their lack of willingness for SA their work are concerns for copyright issues, additional time and effort needed to SA and the age of the researchers (Davis & Conolly, 2007; Kim, 2010).

TAPoR is a project that is still running and the website and data is still accessible online. Technically we cannot simply talk of archiving because the original data is still available in its primary storage location in a format suitable for multiple quick and dynamic access. What we are doing is a combination of archiving and backup but we will continue to use the term archiving to discuss our ‘archiving’ process. We want to encourage SA and the practise of preparing and depostiong the archive (the data itself to be deposited) from the beginning and throughout the lifetime of projects.

Different organisations have different policies of how data is archived. Some regulatory bodies will require that all applications used to access the data (and sometimes even hardware) be preserved with the data while others will require that only the data itself be archived. This relates to the old debate in archiving of emulation versus conversion where emulation refers to the archiving of data in its original format such that the use of the data is emulated in the archive while conversion involves the data being converted into more recent formats for optimal use (Dollar, 1999; Doorn & Tjalsma, 2007). But the main requirements from the archiving process is for the data to be recoverable. The format of storage might not always be the same and it is becoming common to see archives that use XML and JSON to effectively store metadata (Reeve, 2013). As a project is run, how data is collected, used and saved might change. The archiving process has to be able to accommodate these changes and decisions will have to be made along the way to either keep both formats or to choose one and make all data conform to the single format. If the data only needs to be assessed and not to be converted back in the

original environment in which it was used (for example a webpage, or visualization), best practise suggests it might be good to transform the data into a common format (Reeve, 2013) and care has to be taken as metadata are easily lost in the process. The data saved in the archive has to be accessible such that someone trying to access it in 1 year, 10 years, 100 years or more should be able to do so successfully without facing any difficulty. The format used should be a common format which will not become unsupported in the foreseeable future. The Library and Archives Canada (LAC) (2014) of the government of Canada recommends the use of PDF and simple text encoded in Unicode for text data and the used of XML with document type definition (DTD) for the storage of a database file (Library and Archives Canada, 2014). This shows a move away from database management systems to like SQL to NoSQL. The general recommendation is to use non-proprietary, multiplatform and uncompressed formats.

Archiving TAPoR

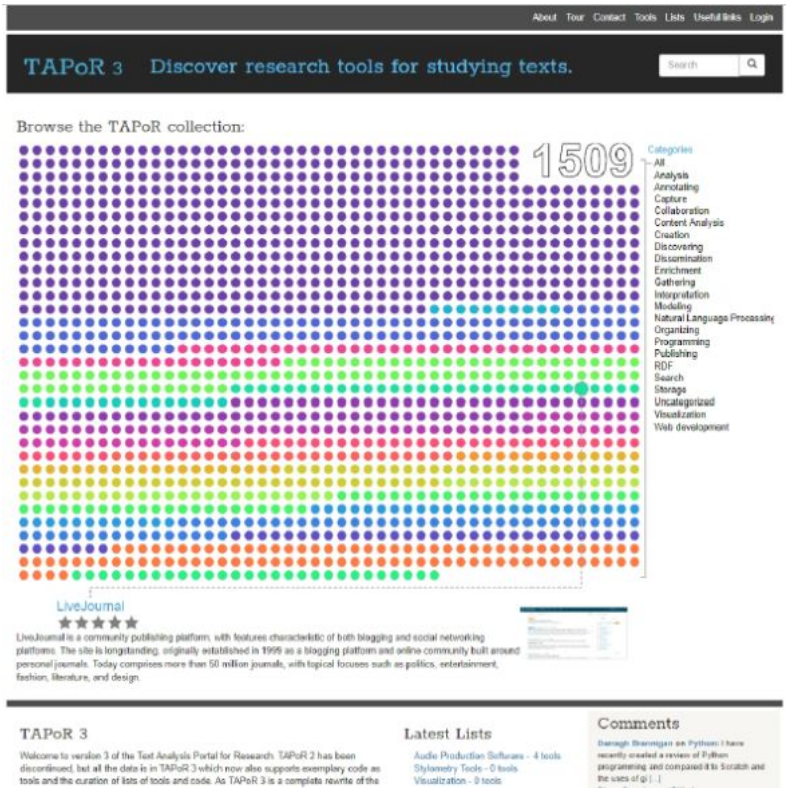
TAPoR is still running; new tools continue to be added; users continue to leave comments on tools and the TAPoR team is still open to suggestions on how to update TAPoR to make it better. It is currently in its third iteration and the TAPoR project has undergone many changes over the years of its existence. This is the first time that an archive of TAPoR was created and deposited. The former iterations of TAPoR—mostly the user interface and functionalities—have been lost, this means the current archive is lacking in that it doesn't document the history of TAPoR, but as the saying goes, better late than never. This is just the first archive and depending on the amount of change at the time TAPoR will be archived again, this archive will be updated or an entirely new one will be created.

Our objective when we started the archiving process was to create an archive with data that was easily accessible to someone in the present day and in 100+ years in the future. We

did not only want to deposit the data but we also wanted to present how the data was accessed through the TAPoR website user interface. As mentioned above one of the reasons researchers don't SA their research projects is the time and effort needed to prepare and deposit the archive. We also wanted to make the process as simple as possible so that other researchers can use it as an example and are encouraged to deposit their projects. We decided that we had to archive not only the data but also the user interface so that archaeologists of the future could see how using TAPoR was. The different documents that made up the archive are presented next.

The user interface

Archiving the user interface was to permit someone accessing the archive in the far future when TAPoR might no longer be running or in a nearer future when the interface and functionalities might have changed to see what using the website as it is today was. Although some archivists advocate for saving the user interface in html format, we opted for just taking screenshots of the main web pages that makeup the TAPoR website. For the html code to display the user interface correctly, the browser or viewer has to support many programming languages, many of which will surely no longer be supported in the next few decades. The screenshots of each webpage were combined into a PDF document—using PDF24 available for free at https://en.pdf24.org/—and placed in a 'User interface' folder. Extracts of the homepage and the Tools page as saved in the archive are shown in figure 1 below.


About Tour Contact Tools Lists Useful links Login
TAPoR 3 Discover research tools for studying texts.
Search
Browse the TAPoR collection:
1509
Categories
All
Analysis
Annotating
Capture
Collaboration
Content Analysis
Creation
Discovering
Dissemination
Enrichment
Gathering
Interpretation
Modeling
Natural Language Processing
Organizing
Programming
Publishing
RDF
Search
Storage
Uncategorized
Visualization
Web development
LiveJournal
LiveJournal is a community publishing platform, with features characteristic of both blogging and social networking platforms. The site is longstanding, originally established in 1999 as a blogging platform and online community built around personal journals. Today comprises more than 50 million journals, with topical focuses such as politics, entertainment, fashion, literature, and design.
TAPoR 3
Welcome to version 3 of the Text Analysis Portal for Research. TAPoR 2 has been discontinued, but all the data is in TAPoR 3 which now also supports exemplary code as tools and the curation of lists of tools and code. As TAPoR 3 is a complete rewrite of the
Latest Lists
Audio Production Software - 4 tools
Stylometry Tools - 0 tools
Visualization - 0 tools
Comments
Darragh Brannigan on Python: I have recently created a review of Python programming and compared it to Scratch and the uses of gi [...]

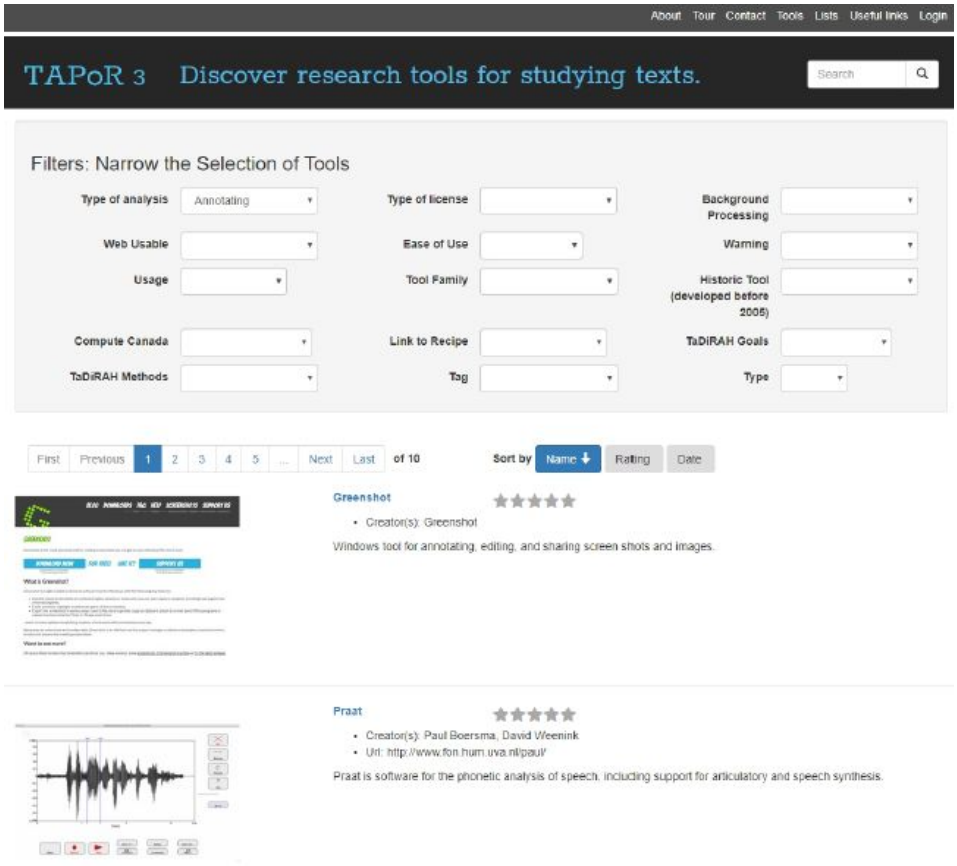


fig 1:(top) screenshot of homepage, (bottom) screenshot of the tools page

The data

TAPoR's data is information about tools that are used not only for text analysis, but that span many categories including: communication tools, repositories and archives, GIS tools, browser extensions, multimedia files editors and development tools. TAPoR being a database driven site, the data was exported from the database as a single dump file. Importing the data into a database management system (DMS) wasn't a straightforward task as we faced many incompatibility issues and bugs in our DMS version. This made it clear that depositing the data

as an SQL dump file was not a good option if we want the data to be accessible to someone 100+ years in the future. It was easy to imagine digital archaeologists 200 years in the future being frustrated after getting an error message when trying to import the dump file despite using a DMS of our time but a version that had some bugs or some incompatibility issues. After importing the dump file into phpMyAdmin we decided to export the data in XML format. Each table of the relational database was converted into an XML document and was saved in the 'XML tables' folder of the archive. Both the dump file and the XML data were included in the archive. With the right DMS, importing the data from the dump file is much easier and manipulating the data in a DMS is also much easier. There were 17 tables in the database. For each XML document created from a table, each data entry (row) had as parent tag 'table' with the name of the table as an attribute. For each entry, the data for each column was placed in a 'column' tag with the name of the column as an attribute. An image of the XML document of the tools table is shown in figure 2 below.

```
<table name="tools">
    <column name="id">8</column>
    <column name="user_id">NULL</column>
    <column name="name">Voyant Cirrus</column>
    <column name="detail">&lt;p&gt;Cirrus is a visualization tool that displays a word cloud re
    <column name="url">http://voyant-tools.org/tool/Cirrus/</column>
    <column name="is_approved">1</column>
    <column name="creators_name">Stéfan Sinclair and Geoffrey Rockwell</column>
    <column name="creators_email">sgs@mcmaster.ca</column>
    <column name="creators_url">http://stefansinclair.name/</column>
    <column name="image_url">images/tools/0/8.png</column>
    <column name="star_average">5</column>
    <column name="is_hidden">0</column>
    <column name="last_updated">2018-10-05</column>
    <column name="documentation_url">http://docs.voyant-tools.org/tools/cirrus/</column>
    <column name="code">NULL</column>
    <column name="repository"></column>
    <column name="language">NULL</column>
    <column name="nature">0</column>
    <column name="created_at">2011-05-26 06:38:13</column>
    <column name="updated_at">2018-10-05 03:23:44</column>
    <column name="recipes"></column>
</table>
<table name="tools">
    <column name="id">9</column>
    <column name="user_id">NULL</column>
    <column name="name">Voyant Links</column>
    <column name="detail">&lt;p&gt;Links finds collocates for words and displays links between t
    <column name="url"></column>
    <column name="is_approved">1</column>
    <column name="creators_name">Stéfan Sinclair and Geoffrey Rockwell</column>
    <column name="creators_email">sgs@mcmaster.ca</column>
    <column name="creators_url">http://stefansinclair.name/</column>
    <column name="image_url">images/tools/0/9.png</column>
    <column name="star_average">3</column>
```

Fig 2: XML version of tools table in database

The file size of the dump file was 3MB but when it was converted to XML format, the total file size was 11.7MB. The data we were dealing with was small compared to some datasets that are in the realm of gigabytes and terabytes. Computers and software are becoming better at handling large text files and some can load a few gigabytes of text. It is common to see computers with 16GB of Ram these days and it is expected that storage technology will increase in the more distant future and RAM size will be measured in terabytes. Text files have a good compression ratio so after downloading the data, the user can store the files in a compressed format.

Table description

A table description text file was included in the archive. It gave a tabular representation of each table in the database, its columns and the characteristics of each column. It was important to include this document in the archive because it contained useful information about how the database was designed.

API documentation

Because TAPoR is still running and new tools are constantly being added the API documentation was included in the archive such that a user could be able to get the most recent information about tools listed in TAPoR directly from the TAPoR database. The role of the archive is to make data available and the possibility to make more recent data available should be provided if possible in the archive. In a recently published paper, a group of researchers who had the SQL dump file chose to use the API to get a more recent list of tools in TAPoR. They investigated the frequency of mentions of the tools in The Alliance of Digital Humanities Organizations (ADHO) conference abstracts to get an idea of the most used tools in the DH field (Barbot, Fischer, Moranville & Pozdniakov, 2019).

The readme document

A readme text file was created that gives brief description of the documents that make up the archive. It has instructions on how to open the SQL dump file which is the only file in the archive that is not in a common format like text. It also contained a brief description of what TAPoR is. This was included because someone not necessarily searching for TAPoR can come across the

archive and should be able to quickly know the content of the archive and the user might find it interesting and explore the archive.

Depositing the archive

The choice of the repository in which to deposit an archive is an important one. Care has to be taken to make sure that the institution or repository is legally committed to maintain the archive forever. Such level commitment is usually held by big institutions like national libraries whereas smaller institutions like research institutes are usually considered less trustworthy for long term preservation because of a higher possibility of loss of funding, change of interest of the researchers or staff (Lecher, 2006). We chose the university of Alberta Education and Research Archive (ERA) as the first place to deposit the archive. ERA (https://era.library.ualberta.ca/) is the University of Alberta's open access digital archive created in 2010. ERA is connected to the university library search engines and the files it keeps are indexed in Google hence files deposited in ERA can be found by simple searches. The archive can be consulted here https://era.library.ualberta.ca/items/78117450-301b-401a-87f9-938900c123ef. We plan to deposit the archive in more institutional repositories in the near future. As TAPoR keeps on being updated, we will do yearly updates of the archive in all the repositories in which it is deposited.

Conclusion

XML database is a category of NOSQL which is becoming more common these days. The self-documenting property of XML—with parent child relationships and attributes—makes it a good choice for archiving documents. XML databases can be queried and exported to

different formats. What XML affords is that it makes the data accessible and simple to understand and the user can in simple steps export the data into a preferred format.

Although we advocate for a simple process for archiving database driven websites—to avoid compatibility issues that will very likely occur if a functional website with all the programing languages, protocols, plugins, etc are archived—, it is a fact that much of the richness of the experience of using the website is not archived when screenshots of the website are taken. A better and simple alternative will be to make a video of the website's interface that shows its usage. This video will then be saved in one of the LAC recommended video formats (MOV, AVI) and hence the use experience which according to some archivists is an essential part of the archive will be documented although not emulated. Some future digital archaeologists will surely be interested in software use in these websites and some might have a research interest or hobby the running of old programs so where possible and if enough resources are available to do so, the software or code behind the website can/should also be deposited. The video of the use interface of the program can be a good reference to understand how the deposited software works. The rate of researchers SA their publication remains low (Gadd & Troll Covey, 2019; Kim, 2010) and it is our hope for the advancement of science that more researchers self-archive not only their publications but also their data and projects. We hope to have shown that making and depositing an archive of a database driven website can be a simple task and the readers are encouraged to plan early and deposit their entire projects which are part of our common digital heritage.

Future directions

SQL databases are very common and efficient and are the usual source file from which XML databases are exported. Exporting an XML database from an SQL database as was done here

using the available DMS does not allow for much customisation. Tables are exported individually and column names are exported as attributes. The ability to create a single XML database document with data extracted from multiple tables and used as either attributes or child tags will be a good addition to DMS. This can also be achieved by writing specific scripts for the data. Alternatively, an application can be created which will analyse the database and provide customization options on what data from the tables should be exported and how they should be exported. Creating this application is one of our future projects.

Reference

Barbot, Fischer, Moranville & Pozdniakov, 2019. Which DH Tools Are Actually Used in Research? weltliteratur.net:A Black Market for the Digital Humanities Retrieved from https://weltliteratur.net/dh-tools-used-in-research/

Brody, T., Carr, L., Gingras, Y., Hajjem, C., Harnad, S., & Swan, A. (2007). Incentivizing the open access research web: publication-archiving, data-archiving and scientometrics. CTWatch quarterly, 3(3).

Brown, S., Clements, P., Grundy, I., Ruecker, S., Antoniuk, J., & Balazs, S. (2009). Published yet never done: The tension between projection and completion in digital humanities research. Available at http://www.digitalhumanities.org/dhq/vol/3/2/000040/000040.html

Davis, P. M., & Connolly, M. J. (2007). Institutional repositories: evaluating the reasons for non-use of Cornell University's installation of DSpace. D-Lib Magazine, 13(3/4).

Dollar CM (1999) Authentic electronic records: strategies for long-term access. Cohasset Associates, Chicago

Doorn, P., & Tjalsma, H. (2007). Introduction: archiving research data. Archival science, 7(1), 1-20.

Gadd, E., & Troll Covey, D. (2019). What does 'green'open access mean? Tracking twelve years of changes to journal publisher self-archiving policies. Journal of Librarianship and Information Science, 51(1), 106-122.

Grant, K., Dombrowski, Q., Ranaweera, K., Rodriquez-Arenas, O., Sinclair, S. and G. Rockwell. (2020). "Absorbing DiRT: Tool Discovery in the Digital Age." *Digital Studies/le Champ Numérique*. Forthcoming.

Kim, J. (2010). Faculty self-archiving: Motivations and barriers. Journal of the American Society for Information Science and Technology, 61(9), 1909-1922.

Kitchin, R. (2014). The data revolution: Big data, open data, data infrastructures and their consequences. Sage.

Lecher, H. E. (2006). Small scale academic web archiving: DACHS. In Web Archiving (pp. 213-225). Springer, Berlin, Heidelberg.

Library and Archives Canada. (2014). Guidelines on File Formats for Transferring Information Resources of Enduring Value available at https://www.bac-lac.gc.ca/eng/services/government-information-resources/guidelines/Documents/file-formats-irev.pdf

Masanès, J. (2005). Web archiving methods and approaches: A comparative study. Library trends, 54(1), 72-90.

Reeve, A. (2013). Archiving data Retrieved from https://www.sciencedirect.com/topics/computer-science/archiving-data

Rockwell, G. (2006). “TAPoR: Building a Portal for Text Analysis”, in *Mind Technologies: Humanities Computing and the Canadian Academic Community*. Ed. Raymond Siemens and David Moorman. Calgary: University of Calgary Press, p. 285-299.

Rockwell, G. and S. Sinclair (2016). *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. Cambridge, Massachusetts, MIT Press.