Improving Collection Understanding in Web Archives

Extended Abstract for the JCDL 2018 PhD Consortium

Shawn M. Jones
Old Dominion University
Norfolk, Virginia
sjone@cs.odu.edu

ABSTRACT

Ever since the Internet Archive started large-scale web archiving in 1996, historians, sociologists, and journalists have found web archives to be an important source of information for their work. Archive-It, a service focused on creating collections, allows curators to generate their own web archive collections. Many of these collections are vast, consisting of thousands of documents. This makes collection understanding a difficult, if not impossible, task. We seek to improve collection understanding by summarizing these collections and visualizing the summaries. Focusing on Archive-It, we seek to identify the different types of web archive collections, the algorithms that can be used summarize those collections, and the best visualizations of those summaries to support better collection understanding.

ACM Reference Format:
Shawn M. Jones. 2018. Improving Collection Understanding in Web Archives: Extended Abstract for the JCDL 2018 PhD Consortium. In *Proceedings of ACM-IEEE Joint Conference on Digital Libraries Doctoral Consortium (JCDL'18).* ACM, New York, NY, USA, 9 pages.

1 INTRODUCTION

With web archives, journalists find evidence and information to back up their stories [48], historians store information for later users [44], and social scientists can study the actions of humans during specific time periods [17]. These different groups gain value not only from creating their own collections, but from using the collections of others. Web archive collections store the content that would otherwise be lost. As users, we currently have no good way of understanding what is in each collection without manually reviewing all of its items. Web archives intentionally consist of different versions of the same document. With these multiple versions we can watch the evolution of a single resource over time, following the changes to an organization or how the public learns the details of an unfolding news story. As aggregations of archived web pages, or **mementos**, these collections become resources unto themselves. While past work has used mementos for studying how web resources change over time [28] or evaluated the changes to various industries [21], there is still serious theoretical work to be done in improving the usability of web archive collections. Our goal is to help collection creators and the public at large to make better use of these collections through improvements to collection understanding.

Archive-It is a popular web archive collection platform developed by the Internet Archive in 2005 [40]. Using Archive-It as a basis for study, we seek to develop methods that would address the problem of collection understanding. With Archive-It, a curator chooses what live web content to preserve in the form of original resources, also known as **seeds**. Each Archive-It collection page shows these seeds in alphabetical order along with the times that they were archived. Both collection pages shown in Figure 1 are about the South Louisiana Flood of 2016. The left collection has 17 seeds, and the one on the right has 68. Each seed has many mementos. How does their content differ? For a researcher studying this event, which collection should they choose? The researcher can evaluate the mementos for each seed, but some collections have thousands of seeds. The Archive-It collection "Government of Canada Publications" has 132,599 seeds, most crawled more than once. In addition, there are often multiple collections about the same topic. The search engine for Archive-It returns a list of 31 collections in response to the query "human rights". With so many seeds and mementos to review, how does a researcher understand the difference between them? Archive-It does allow curators to supply metadata for collections, as shown in Figure 2. But, after a review of public Archive-It collections in 2017, we discovered that, in spite of having access to the metadata fields of Dublin Core [7], metadata is inconsistently applied to collections, likely due to differences in content standards [57] and rules of interpretation among curators. In addition to the issue of scale within a collection, more Archive-It collections are added each year, as shown in Figure 3, reaching more than 8000 collections in 2016. Retroactively applying metadata to collections, seeds, and mementos would be costly in terms of time and personnel. Thus, the metadata is insufficient, and the size of these collections also makes it an expensive proposition for a potential user to review them completely.

To provide assistance to researchers, we seek to automatically generate summaries of collections, similar to Luhn's automatic generation of abstracts [38]. We differ by surfacing the best mementos, rather than just sentences, as elements for our summarization. To borrow concepts from information foraging theory [51], our goal is to maximize the value of the knowledge gained from our collection summary while minimizing the cost of interacting with the collection [50]. Users rely upon textual and visual clues to determine if a resource fits their needs. A resource with textual and visual clues indicating that it meets a user's needs is said to have good **information scent**. Thus, if the collection meets a particular user's needs, we want to ensure that the mementos chosen for our summarization have good information scent .

2 BACKGROUND AND RELATED WORK

To build a web archive collection at Archive-It, a curator selects seeds and submits each seed's URI to a crawler, which then creates mementos based on these seeds at various intervals chosen by the curator. We refer to the mementos created directly from seeds as

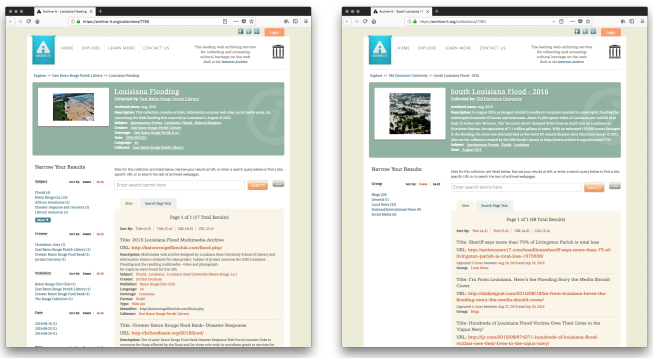

Figure 1: Archive-It collection 7755 (left) and collection 7760 (right) both cover the South Louisiana Flood of 2016.

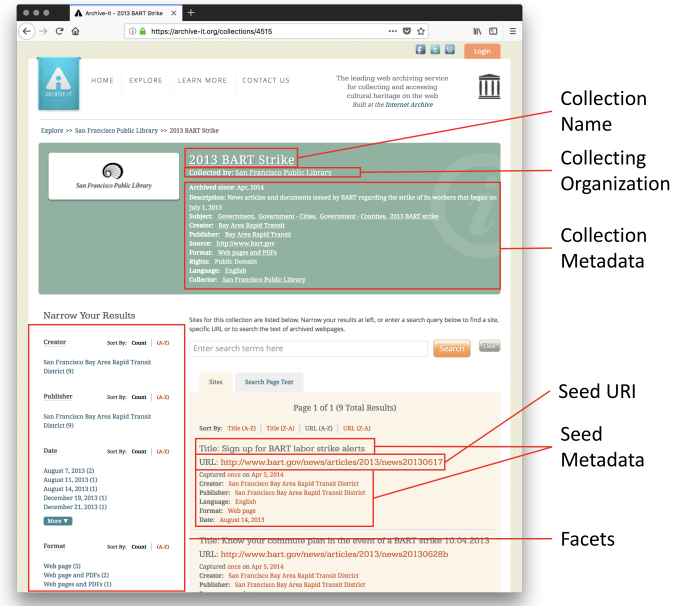

Figure 2: Annotated page for Archive-It collection 4515

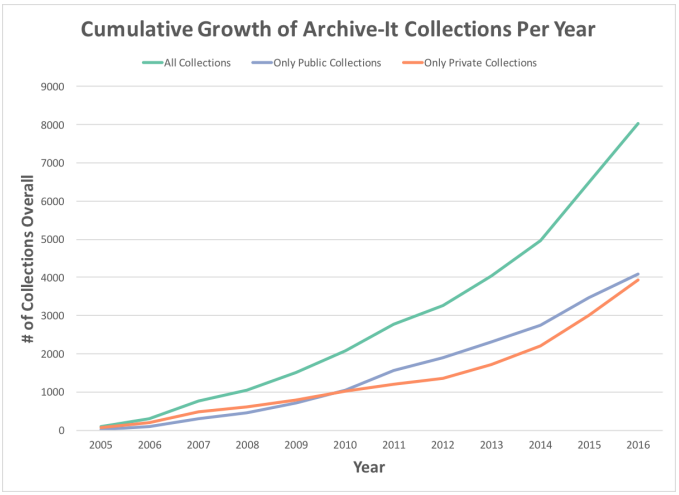

Figure 3: Cumulative Archive-It collection growth

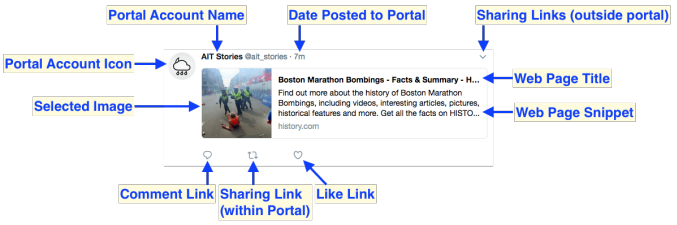

Figure 4: A social card from Twitter for URI https://www.history.com/topics/boston-marathon-bombings, annotated to indicate its parts

seed mementos. In addition, the curator can instruct the crawling software to create mementos of pages linked to from those seeds, and, potentially, pages linked from those pages. The **deep mementos** created via this process were not chosen directly by the curator. Because crawling software uses rules to determine how far to crawl, it is possible that many of these deep mementos were not intended to be part of the collection. The seeds and seed mementos were specifically chosen by the curator and hence they are the focus of our summarization work because they represent unique policy and behavior for each collection. Each seed has an associated machine-readable **TimeMap** listing all of that seed's mementos and the dates and times that they were crawled, their **memento-datetimes** [60].

Understanding the difference between two items, whether they be collections or research papers, is usually handled by a record containing metadata about the publication. The fields for this record are typically encoded in some standard, such as EAD [47]. As noted above, Archive-It collections also allow the curator to supply metadata, both at the collection and the seed level. Unfortunately, the metadata is supplied by collection curators with no consistent, agreed upon rules, resulting in the inability to reliably compare the metadata of one collection with another. The challenge in using such metadata solutions is scale. They were designed with the intention that humans would carefully review each item in the collection and use some cataloging standard to generate that item's metadata. With thousands of documents to review in a web archive collection, this becomes a costly endeavor.

Zhang [63] and Li [35] both developed algorithms for generating summaries from a corpus of documents in order to surface information pertinent to certain **aspects** about events. For example, from a set of news stories about a disaster, a series of sentences are generated addressing concepts like *time*, *place*, *cause*, and *countermeasures*. Their algorithms rank sentences for extraction from these stories so that they can then be included in a summary that addresses these aspects. These aspects correspond to the questions users have when trying to compare items. Rather than ranking sentences, we can rank the mementos in the collection based on how well they address various aspects, thus ensuring that the mementos in our summary have high information scent.

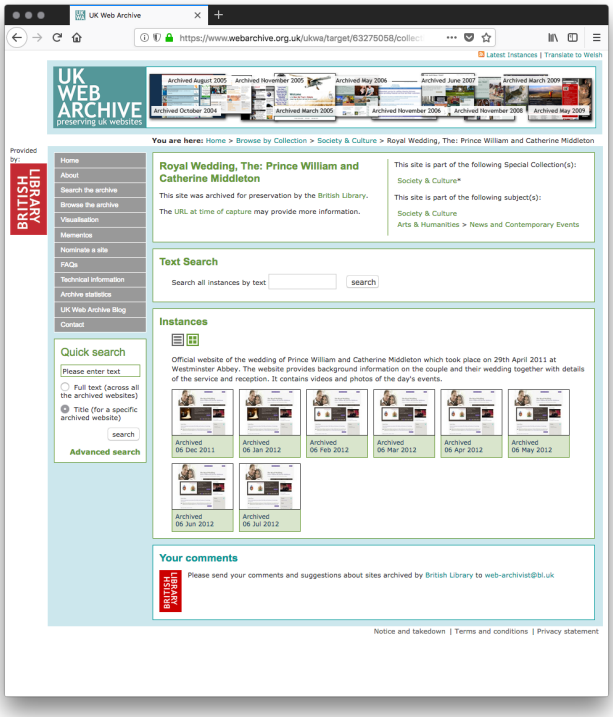

Figure 5: The UK Web Archive uses thumbnails to visualize mementos in its web archive collections.

There are techniques other than sentence extraction for producing good information scent from a corpus. The *septem circumstantiae* attributed to Aristotle, also known as the "five W's (and one H)" of journalism [55], can provide us with handles to convey aspects of aboutness from a corpus. By employing techniques like Named Entity Recognition [46] we can expose mementos that indicate who is mentioned or where the concepts in the collection take place. To answer questions of the time periods mentioned in the collection or even the collection's publication dates, we can employ natural language processing to expose temporal expressions [34]. For a more general comparison of concepts, there exist topic modeling techniques like Latent Dirichlet Allocation [11] and Latent Semantic Analysis [18]. Some work has been done to use these concepts with web archive collections. Sağlam sought to use the content of specific Archive-It collections to analyze the timeliness of medical data through the use of information retrieval techniques [54]. Milligan seeks to understand the past by studying web archive collections [42], recently studying the social aspects of GeoCities [43]. Success has also been found with other types of document collections. Ramage studies how topic modeling can meet the needs of social scientists hoping to understand large corpora of text [52], leading to the development of Stanford's Topic Modeling Toolkit. Chang provides metrics to help evaluate topic models [16]. Blei suggests that user interfaces and visualizations built upon topic modeling would be a successful method of exploring collections [10].

Because web archive collections have a temporal component, there are challenges to using these text analysis methods, as discussed by the Longitudinal Analytics on Web Archive Data (LAWA) project [61]. Named entities change over time and need to be disambiguated. For example, if the entity *Clinton* is mentioned in a memento, is the document referring to, Bill, the US president from the 1990s or to Hillary, the US presidential candidate from 2016? For topic modeling or natural language processing, an algorithm that is not temporally aware may group words together that have different meanings due to changes in language over time. Though LAWA focused on supporting searching of archives, the issues and solutions uncovered by the project will inform how we process web archive collections for summarization.

We want to understand which techniques are best for visualizing our mementos. Loumakis showed that text provides better information scent than images when summarizing web pages, but the scent of text combined with images is higher [37]. Text and images are usually joined together in social cards, with Figure 4 being an annotated example from Twitter. Social cards are now widely encountered and easily understood by web users, making them a good choice for visualizing mementos. Groups of social cards work as a visualization because the social card for a given platform stores the same information in the same place on each card. All Twitter social cards are bounded in a grey box with an image on the left, with the right consisting of a title on the top with a snippet on the bottom. Because of this consistency, the use of multiple social cards on a web page is analogous to the small multiples visualization technique catalogued by Tufte [59]. The presence of the same information field in the same location on each card allows for quick comparisons and hence improves understanding of the whole group.

Thumbnails, screenshots of a memento rendered in a browser, are a common method of rendering mementos [5], as seen in Figure 5. Jiao [23] discovered that internal images pulled from the page itself, like those used in social cards, summarize a given web page better than thumbnails if those images are dominant [36]. On the other hand, thumbnails are preferred for web pages that have a simple structure with clear text and images, even though the thumbnail often contains the dominant image. Because these images were being considered outside of their page, it is assumed that users liked the additional data provided by the thumbnail in these cases. Thumbnails were not preferred by users when the thumbnail contained mostly text. Thus, when summarizing individual mementos, we want to know: when are thumbnails better for memento understanding than social cards?

A social card or a thumbnail is an example of a **surrogate** — a visualization of a single web resource. Surrogates have been of interest to those seeking to improve search result interfaces. Different surrogate types have been compared by Woodruff [62], Dziadosz [19], Li [36], Teevan [58], Aula [8], Al Maqbali [1], Loumakis [37], and Capra [15]. As to the question of which surrogate is best for web resources, Woodruff, Dziadosz, and Teevan came to different conclusions for text compared to thumbnails. Capra, Al Maqbali, and Loumakis came to different conclusions as to whether text or social cards performed better. All of these studies focused primarily on how well participants determined search engine result relevance,

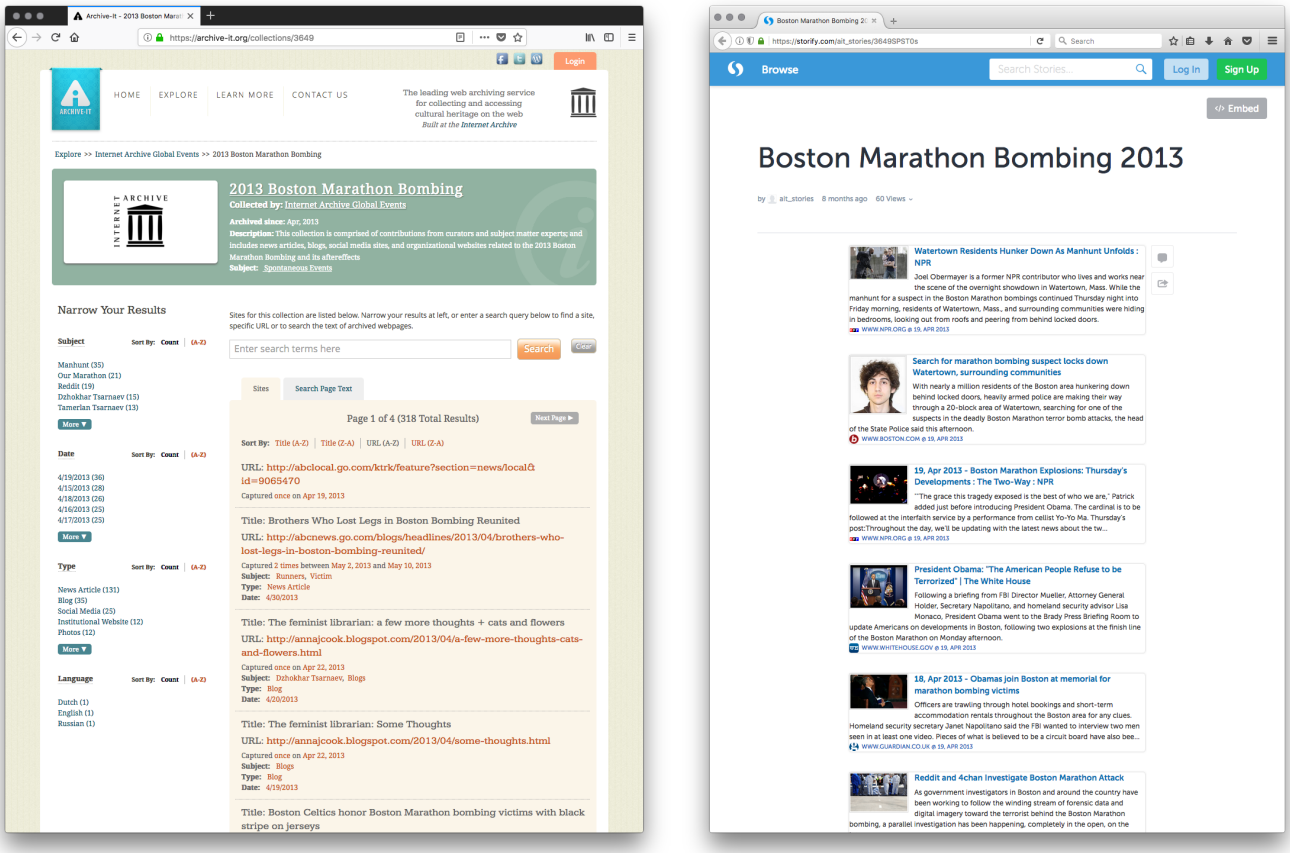

Figure 6: Archive-It collection 3649 about the Boston Marathon Bombing (left) and an example visualization produced by AlNoamany's algorithm, visualized in Storify (right)

but did not compare surrogates for the purpose of collection understanding. In this context, we still believe that the best surrogate for use with mementos from a web archive collection is an open question [25].

Padia [49] tried to characterize each collection as a whole using different visualizations. He employed techniques such as treemaps, wordles, bubble charts, and timelines, providing insight into different aspects of the collection. Because these visualizations tried to include everything in the collection, they had a high degree of visual complexity. Our work seeks to reduce the visual complexity by creating a view into the collection [45], by reducing the number of items for use in a visualization to those that best represent the collection as a whole.

Prior work by AlNoamany [2, 4] shows some success with combining the summarization of web archives with the visualization capabilities of **live web** curation platforms, like Storify and Pinterest. Live web curation platforms store URIs and metadata, but not the documents themselves. Once the documents vanish, only broken links remain. Where web archive collections may contain thousands of items, live web collections often contain less than 30 [3]. Live web curation platforms also provide nice visualizations of individual items usually with social cards. The familiarity with the social card makes live web collections easier to interpret for the user. AlNoamany's prior work on web archive collection understanding summarized collections by selecting a small number of high quality mementos to represent the collection. These **representative mementos** would then be visualized using the storytelling tool Storify, as shown in Figure 6. While this work was significant, AlNoamany primarily focused on web archive collections that centered on spontaneous events, such as The Boston Marathon Bombing of 2013, the 2013 Government Shutdown, and the Wikileaks Document Release of 2010, and not other types of collections such as when organizations archive themselves, collections centered on a particular subject other than an event, or collections covering expected events like the Olympics. She tailored the solution to be visualized in Storify, which will cease operations in May of 2018 [56], causing us to search for potential replacements [24]. AlNoamany proved that test participants could not tell the difference between stories created by a human or created via this algorithm. She did not, however, evaluate whether or not the stories produced were fit for the purpose of collection understanding.

The algorithm can be summarized in the following steps:

(1) Acquire content of seed mementos.

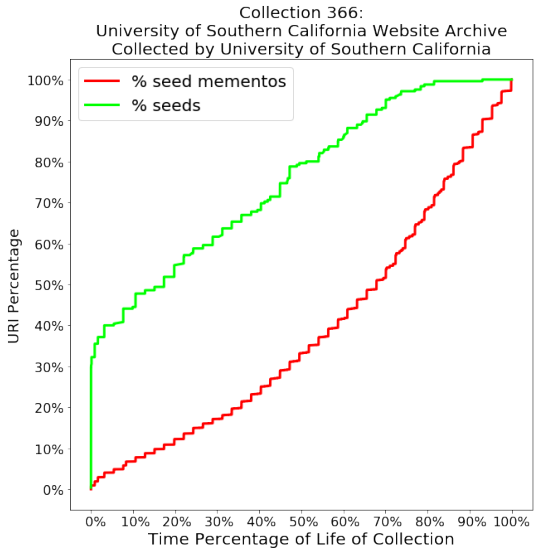

Figure 7: Example growth curve for Archive-It collection 366

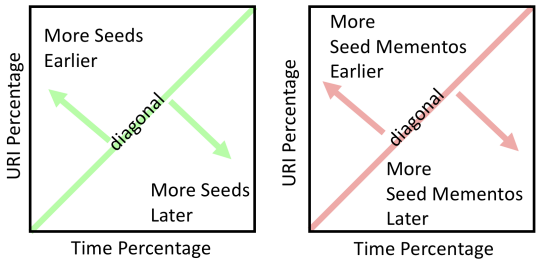

Figure 8: Anatomy of an Archive-It collection growth curve

(2) Remove seed mementos in the collection that are off-topic.
(3) Remove non-English seed mementos.
(4) Remove duplicate seed mementos.
(5) Slice the collection into k buckets based on memento-datetime to ensure the summary spreads across the time period of the collection.
(6) Cluster each slice based on content similarity to discover novel seed mementos that can satisfy a large range of information needs among users.
(7) Find the best seed memento from each cluster based on their memento-damage [14], URI depth [41], and potential for high quality metadata [49].
(8) Order the seed mementos by memento-datetime
(9) Visualize these seed mementos with social media storytelling

Steps 5 - 9 expect the Archive-It collection to be centered on an event in order to work well and expect Storify to be the target of the visualization. The memento selection is informed by memento-datetime and novelty, rather than aspects of the collection. We seek to prove that new algorithms and visualizations informed by the aspects and types of collections support collection understanding.

3 PRELIMINARY WORK

Even though AlNoamany had studied some features of Archive-It [3], we wanted a better understanding of individual collections. As part of a review of 3,382 public Archive-It collections [27], we identified several structural features that provide different dimensions of insight into the curatorial behavior of each. In Figure 7 we show a **collection growth curve** for Archive-It collection 366, borrowed from work about the growth of entire archives by Al-Sum [6]. The x-axis represents the **life of the collection**, or time between the first memento and the last. The y-axis represents the percentage of URIs in the collection at a given time. The URIs come in two categories: seeds, represented by green, and seed mementos, represented by red. Figure 8 shows the relationship between the curve and the diagonal, providing insight into the behavior of the curator. This behavior can be quantified by using the area under the curve (AUC) of each of these curves.

By taking the **difference between the seed curve AUC and diagonal**, we gain insight into when seeds were added during the collection's life. If this value is positive, then more seeds were added earlier, and the curve bends to the upper right. If this value is negative, then more seeds were added later. Similarly, we can use the **difference between the seed memento curve AUC and the diagonal** to study the crawling behavior and hence the growth of the actual collection. If this value is positive, with most of the mementos existing in the upper left, then most of the collection was filled earlier in its life, and opposite if the value is negative. The **difference between the seed curve AUC and seed memento curve AUC** indicates the distance between when the curatorial decisions were made and when the mementos were added to the collection.

In addition to the curves, there exist other features useful for understanding collections. The **number of seeds** submitted to the collection varies, as does the **number of seed mementos**. These can be counted by acquiring seed lists for each collection and analyzing their TimeMaps. The **lifespan** of the collection can be calculated by taking the time difference between the collection's first and last memento.

The seed URIs for each collection contain their own insight. **Seed URI domain diversity** quantifies the spread of the collection across different sources. A collection where all seeds are from the same domain would have a domain diversity of 0 and one where all seeds are from different domains would have a domain diversity of 1. The **path depth** for each seed URI consists of the number of items separated by slashes after the domain name [41]. We acquire an idea of the spread of path depth across the collection with the **seed URI path depth diversity** metric. This may indicate if the seed URIs consist solely of top-level pages or a mixture of top-level pages and more specific content. A different metric, **most frequent URI path depth**, can be used to understand the actual path depth values. If this value is 0, then the collection's seeds are mostly top-level pages of web sites. If the most frequent path depth is higher, then it mostly consists of seeds deeper in a web site. Some collections consist mostly of seed URIs with query strings, whereas others consist of just paths. We capture this behavior with the **% query string usage** metric.

Table 1: Distribution of collections for each semantic category

Semantic Category	# of Collections	% of All Collections
Self-Archiving	1,828	54.1%
Subject-based Archiving	935	27.6%
Time Bounded - Expected	476	14.1%
Time Bounded - Spontaneous	143	4.2%
Total	3,382	100%

All of these features have a role to play in our eventual collection summarization techniques. Collections with their curves skewed in a given direction may require more focus on the mementos from that time period of the collection. Collections with a few seeds and few mementos might be handled with simpler logic and less processing than more populous collections. Those with high domain diversity and high path depth diversity may be optimized for novelty.

We reviewed the Archive-It collections by hand and placed them into four **semantic categories**.

Self-Archiving — These collections consist of one or more domains either (1) belonging to the archiving organization, or (2) being archived as part of some archiving initiative of which the collecting agency is part. Collections fitting into this category include the *University of Utah Web Archive*, or the *City of Eagan Websites*.

Subject-based Archiving — Some collections consist of a number of seeds bound by a single topic, such as *Environmental Justice*, archived by Tufts University.

Time Bounded - Expected — These collections focus on an expected, planned event, such as *2008 Olympics* archived by the University of North Carolina. The collections may also be based on a specific time period, such as *Virginia's Political Landscape, 2007* archived by the Library of Virginia.

Time Bounded - Spontaneous — These collections start after a spontaneous event. Collections fitting into this category include *Tucson Shootings* archived by the Virginia Tech: Crisis, Tragedy, and Recovery Network and *2011 Japan Earthquake* archived by the University of Michigan, School of Information.

Using Weka v. 3.8.2 [20] with the semantic category set as the target class, we determined that, with $F_1 = 0.720$, Random Forest [13] is the best classifier for predicting the semantic category from the structural features above via 10-fold cross validation.

The distribution of these semantic categories is shown in Table 1. As noted before, AlNoamany only focused on collections centered on events, a category we named *Time Bounded - Spontaneous*, which turns out to be the smallest category. Predicting these categories with Random Forest is a first step. We intend to study the other categories in more detail to understand how best to summarize them.

4 RESEARCH PLAN

The research plan for this work is shown in Figure 9. We are in the process of developing a framework inspired by AlNoamany's algorithm, under the project name Dark and Stormy Archives (DSA). The goal is to develop a flexible framework where individual parts

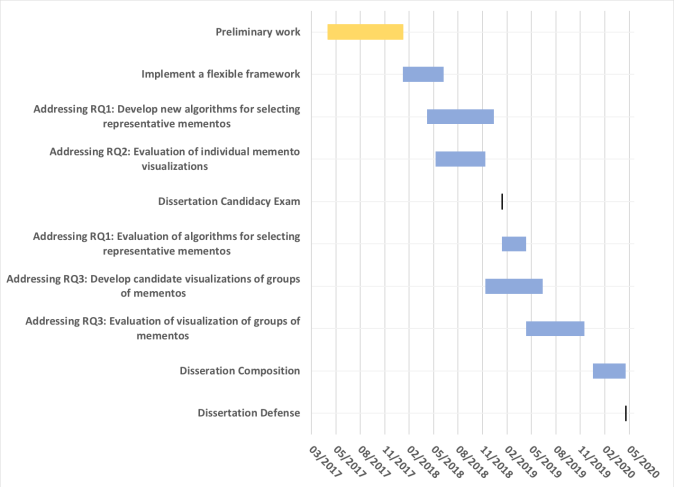

Figure 9: Gantt chart of research plan

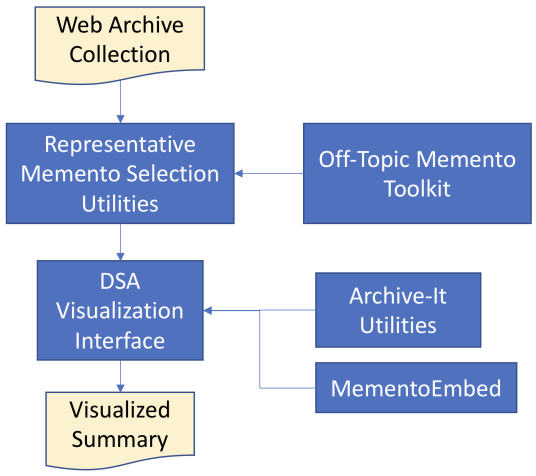

Figure 10: Architecture of the system to be developed for this research

of the algorithm can be swapped out and exchanged, as shown in Figure 10. Our work consists of two concepts: selecting representative mementos and visualizing those mementos. The **Representative Memento Selection Utilities** will provide a common interface for users to select the best representative mementos from a collection. It will make use of the **Off-Topic Memento Toolkit**[1] to filter mementos that are off-topic or otherwise unsuitable to be included in further processing [29]. The **DSA Visualization Interface** will provide a common interface for interacting with different live web curation platforms. It will use **MementoEmbed**[2] to generate surrogates, such as social cards and thumbnails [26]. It will use **Archive-It Utilities**[3] for extracting information directly from Archive-It collections.

[1] https://github.com/oduwsdl/off-topic-memento-toolkit
[2] https://github.com/oduwsdl/MementoEmbed
[3] https://github.com/oduwsdl/archiveit_utilities

To evaluate how well our summarizations and visualizations perform, we will need to choose target collections for study and manually develop **user tasks** [30] for each collection. These user tasks will consist of being able to understand one or more aspects of the collection, whether it is the entities represented in *Self-Archiving* collections, or the events in an unfolding news story in *Time Based - Spontaneous* collections.

With the knowledge of different metrics and semantic types of Archive-It collections, we seek to address the following research questions.

RQ1: How do we select representative mementos for the different semantic types of collections?

Summarization of a single Archive-It collection requires grouping the mementos in the collection by their commonalities, and then selecting the highest quality mementos from each group. We intend to explore many different techniques for grouping and then discovering representative mementos with strong information scent. To do so, we will research algorithms that use techniques that focus on surfacing specific aspects of the collection. Depending on what needs to be surfaced, we may need to apply different techniques to different parts of a collection. We intend to focus on solutions that make use of existing off-the-shelf products where possible, such as spaCy [22], gensim [53], and Stanford CoreNLP [39].

This, of course, does not mean that we will stop at merely selecting mementos. If there is metadata, such as entities or topics, gathered during the process of selecting these mementos that might be useful as part of any further visualization, we will surface that as well.

Evaluation. Questions to be addressed for each algorithm and semantic category of a collection:

(1) How many user tasks were addressed by the mementos chosen? How many user tasks failed?
(2) How many mementos produced are not useful for any user task?
(3) Which algorithm surfaces aspects satisfying the highest mean number of user tasks for a given collection type?
(4) What is the mean minimum number of mementos necessary to address the most user tasks?

Because these questions require analyzing the complete content of multiple mementos, the evaluation of these questions will be done via automated text analysis based on our knowledge of the collections. Through this testing we expect to eliminate several candidate algorithms before we even consider their output for visualization.

RQ2: What surrogate works best for understanding individual mementos?

Because past work was influenced by Storify, visualization elements were fixed. We have an opportunity to study which surrogates, social cards or thumbnails, work best for understanding individual mementos. This work can be done in parallel with the algorithm development and evaluation.

Evaluation. Questions to be addressed for individual memento visualization:

(1) Does the depth, domain, or category of the URI play a factor in which visualization is better?
(2) Do different visualization elements work better for different semantic types?
(3) For social cards, which elements of the social card need to be present to understand the underlying memento?
(4) For thumbnails, what size thumbnail works best for understanding? How much of the web page needs to be rendered for a thumbnail to be useful for understanding?

In these cases, we will engage in user testing, likely with a service like Amazon's Mechanical Turk (MT). When conducted with the MT environment in mind, MT has been shown to provide results comparable to in-person surveys [9] and can even be used for complex tasks [32]. Understanding can be evaluated via user tasks informed by Anderson and Krathwohl's later revision [33] of Bloom's taxonomy [12]. The later revision has been successfully used by Kelly for evaluating participant performance in search tasks [31].

Each participant will be given various user tasks specific to the subset of mementos presented to them gathered from different semantic types of collections. Through this testing, we hope to understand which visualization technique to use for individual mementos prior to creating the small multiples visualization of many of them. As we conduct this study, we will be cognizant of issues with user studies such as fatigue and learning [30].

RQ3: How well do visualizations of groups of mementos produced by different summarization algorithms work for collection understanding?

Once we have determined which individual surrogates work best for a given task and semantic type of collection, we can then evaluate how well they work in combination to allow users to complete a task. There are many different types of visualizations to be generated. Again, freed from Storify, we do not need to restrict ourselves to the "downward flow" visualization of social cards presented in the order of their memento's publication. We could, for example, list thumbnails under a headings containing the name of entities within those thumbnails. We might also produce a visualization consisting of additional metadata gathered by the different algorithms that produced the summary. To limit the cognitive load on participants, we want to focus first on existing visualization paradigms, such as those supported by Twitter Moments and Facebook.

We intend to develop several visualization candidates influenced by the different types of semantic collections, the algorithms discovered as part of answering Research Question 1, and the results from evaluating Research Question 2.

Evaluation. Questions to be addressed for evaluating the visualization as a whole:

(1) How many user tasks are addressed by the visualization chosen? How many fail?
(2) How many visualized mementos were not needed for any given user task?

(3) Given an aspect of the collection, can the user address a user task concerning it by visually scanning the visualization?
(4) Given multiple aspects of the collection, can the user successfully compare different individual memento visualizations to address a user task?
(5) Which visualizations work better for certain semantic types?

Much like with Research Question 2, we will engage in user testing. Each participant will be given various user tasks to specific to the collection to perform. This will be the final set of research and evaluation done on this research project, hoping to address which summarization algorithms combined with which visualization techniques work best for collection understanding.

5 CONCLUSION

By the end of this work, we will contribute more knowledge on the use of collection understanding in web archive collections. We have identified different semantic categories of web archive collections at Archive-It. We have also shown that the semantic categories can be predicted using the structural features of each collection. The existing summarization algorithm for Archive-It collections focuses on only one of these semantic categories, and the smallest one at that. By the end of this research, we will have developed and evaluated new algorithms and visualizations for use in collection understanding of many types of web archive collections. This work will serve as a unique contribution, not because it will merely visualize web archive collections, but because it will provide summarizations of web archive collections using different reduction techniques and well known visualization paradigms.

REFERENCES

[1] Hilal Al Maqbali, Falk Scholer, James Thom, and Mingfang Wu. 2010. Evaluating the Effectiveness of Visual Summaries for Web Search. In *Proceedings of the 15th Australasian Document Computing Symposium*. Melbourne, Australia, 8. http://www.cs.rmit.edu.au/adcs2010/proceedings/pdf/paper%2013.pdf

[2] Yasmin AlNoamany. 2016. *Using Web Archives to Enrich the Live Web Experience Through Storytelling*. Ph.D. Dissertation. Old Dominion University.

[3] Yasmin AlNoamany, Michele C. Weigle, and Michael L. Nelson. 2016. Characteristics of social media stories. *International Journal on Digital Libraries* 17, 3 (01 Sep 2016), 239–256. https://doi.org/10.1007/s00799-016-0185-3

[4] Yasmin AlNoamany, Michele C. Weigle, and Michael L. Nelson. 2017. Generating Stories From Archived Collections. In *Proceedings of the 2017 ACM on Web Science Conference (WebSci '17)*. ACM, Troy, New York, USA, 309–318. https://doi.org/10.1145/3091478.3091508

[5] Ahmed Alsum and Michael L. Nelson. 2014. Thumbnail Summarization Techniques for Web Archives. In *Proceedings of the 36th European Conference on Information Retrieval*. Amsterdam, Netherlands.

[6] Ahmed AlSum, Michele C. Weigle, Michael L. Nelson, and Herbert Van de Sompel. 2014. Profiling web archive coverage for top-level domain and content language. *International Journal on Digital Libraries* 14, 3 (2014), 149–166. https://doi.org/10.1007/s00799-014-0118-y

[7] Ann Apps. 2013. Guidelines for Encoding Bibliographic Citation Information in Dublin Core Metadata. http://dublincore.org/documents/dc-citation-guidelines/.

[8] Anne Aula, Rehan M. Khan, Zhiwei Guan, Paul Fontes, and Peter Hong. 2010. A comparison of visual and textual page previews in judging the helpfulness of web pages. In *Proceedings of the 19th international conference on World wide web*. ACM Press, Raleigh, North Carolina, USA, 51–60. https://doi.org/10.1145/1772690.1772697

[9] Christoph Bartneck, Andreas Duenser, Elena Moltchanova, and Karolina Zawieska. 2015. Comparing the Similarity of Responses Received from Studies in Amazon's Mechanical Turk to Studies Conducted Online and with Direct Recruitment. *PLOS ONE* 10, 4 (04 2015), 1–23. https://doi.org/10.1371/journal.pone.0121595

[10] David M. Blei. 2012. Probabilistic Topic Models. *Commun. ACM* 55, 4 (April 2012), 77–84. https://doi.org/10.1145/2133806.2133826

[11] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (March 2003), 993–1022. http://dl.acm.org/citation.cfm?id=944919.944937

[12] Benjamin S. Bloom. 1956. *Taxonomy of Educational Objectives, Handbook 1: Cognitive Domain*. Addison-Wesley Longman Ltd, Boston, MA.

[13] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (01 Oct 2001), 5–32. https://doi.org/10.1023/A:1010933404324

[14] Justin F. Brunelle, Mat Kelly, Hany SalahEldeen, Michele C. Weigle, and Michael L. Nelson. 2015. Not all mementos are created equal: measuring the impact of missing resources. *International Journal on Digital Libraries* 16, 3 (2015), 283–301. https://doi.org/10.1007/s00799-015-0150-6

[15] Robert Capra, Jaime Arguello, and Falk Scholer. 2013. Augmenting web search surrogates with images. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM Press, San Francisco, California, USA, 399–408. https://doi.org/10.1145/2505515.2505714

[16] Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *Proceedings of the 22Nd International Conference on Neural Information Processing Systems (NIPS'09)*. Curran Associates Inc., USA, 288–296. http://dl.acm.org/citation.cfm?id=2984093.2984126

[17] Renata Gonçalves Curty and Ping Zhang. 2011. Social commerce: Looking back and forward. *Proceedings of the American Society for Information Science and Technology* 48, 1 (2011), 1–10. https://doi.org/10.1002/meet.2011.14504801096

[18] Susan T. Dumais. 2005. Latent semantic analysis. *Annual Review of Information Science and Technology* 38, 1 (2005), 188–230. https://doi.org/10.1002/aris.1440380105

[19] Susan Dziadosz and Raman Chandrasekar. 2002. Do Thumbnail Previews Help Users Make Better Relevance Decisions about Web Search Results?. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. Tampere, Finland, 365–366. https://doi.org/10.1145/564376.564446

[20] Eibe Frank, Mark A. Hall, and Ian H. Witten. 2016. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"* (fourth ed.). Morgan Kaufmann.

[21] Katie Hafner and Griffin Palmer. 2017. Skin Cancers Rise, Along With Questionable Treatments. https://www.nytimes.com/2017/11/20/health/dermatology-skin-cancer.html. *The New York Times* (November 2017).

[22] Matthew Honnibal and Mark Johnson. 2015. An Improved Non-monotonic Transition System for Dependency Parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 1373–1378. https://aclweb.org/anthology/D/D15/D15-1162

[23] Binxing Jiao, Linjun Yang, Jizheng Xu, and Feng Wu. 2010. Visual Summarization of Web Pages. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10)*. ACM, Geneva, Switzerland, 499–506. https://doi.org/10.1145/1835449.1835533

[24] Shawn M. Jones. 2017. 2017-08-11: Where Can We Post Stories Summarizing Web Archive Collections? http://ws-dl.blogspot.com/2017/08/2017-08-11-where-can-we-post-stories.html.

[25] Shawn M. Jones. 2018. 2018-04-24: Let's Get Visual and Examine Web Page Surrogates. http://ws-dl.blogspot.com/2018/04/2018-04-24-lets-get-visual-and-examine.html.

[26] Shawn M. Jones. 2018. 2018-08-01: A Preview of MementoEmbed: Embeddable Surrogates for Archived Web Pages. http://ws-dl.blogspot.com/2018/08/2018-08-01-preview-of-mementoembed.html.

[27] Shawn M. Jones, Alexander Nwala, Michele C. Weigle, and Michael L. Nelson. 2018. The Many Shapes of Archive-It. In *Proceedings of the International Conference on Digital Preservation (iPres)*. Boston, MA.

[28] Shawn M. Jones, Herbert Van de Sompel, Harihar Shankar, Martin Klein, Richard Tobin, and Claire Grover. 2016. Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content. *PLOS ONE* 11, 12 (12 2016), 1–32. https://doi.org/10.1371/journal.pone.0167475

[29] Shawn M. Jones, Michele C. Weigle, and Michael L. Nelson. 2018. The Off-Topic Memento Toolkit. In *Proceedings of the International Conference on Digital Preservation (iPres)*. Boston, MA.

[30] Diane Kelly. 2007. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends® in Information Retrieval* 3, 1–2 (2007), 1–224. https://doi.org/10.1561/1500000012

[31] Diane Kelly, Jaime Arguello, Ashlee Edwards, and Wan-ching Wu. 2015. Development and Evaluation of Search Tasks for IIR Experiments Using a Cognitive Complexity Framework. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval (ICTIR '15)*. ACM, Northampton, Massachusetts, USA, 101–110. https://doi.org/10.1145/2808194.2809465

[32] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, Florence, Italy, 453–456. https://doi.org/10.1145/1357054.1357127

[33] David R. Krathwohl. 2002. A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice* 41, 4 (2002), 212–218. https://doi.org/10.1207/s15430421tip4104_2

[34] Erdal Kuzey, Jannik Strötgen, Vinay Setty, and Gerhard Weikum. 2016. Temponym Tagging: Temporal Scopes for Textual Phrases. In *Proceedings of the 25th*

International Conference Companion on World Wide Web (WWW '16 Companion). International World Wide Web Conferences Steering Committee, Montréal, Québec, Canada, 841–842. https://doi.org/10.1145/2872518.2889289

[35] Peng Li, Yinglin Wang, Wei Gao, and Jing Jiang. 2011. Generating Aspect-oriented Multi-document Summarization with Event-aspect Model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1137–1146. http://dl.acm.org/citation.cfm?id=2145432.2145553

[36] Zhiwei Li, Shuming Shi, and Lei Zhang. 2008. Improving Relevance Judgment of Web Search Results with Image Excerpts. In *Proceedings of the 17th International Conference on World Wide Web (WWW '08)*. ACM, Beijing, China, 21–30. https://doi.org/10.1145/1367497.1367501

[37] Faidon Loumakis, Simone Stumpf, and David Grayson. 2011. This Image Smells Good: Effects of Image Information Scent in Search Engine Results Pages. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM '11)*. ACM, Glasgow, Scotland, UK, 475–484. https://doi.org/10.1145/2063576.2063649

[38] H. P. Luhn. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development* 2, 2 (April 1958), 159–165. https://doi.org/10.1147/rd.22.0159

[39] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. 55–60. https://doi.org/10.3115/v1/P14-5010

[40] Marji McClure. 2006. Archive-It 2: Internet Archive Strives to Ensure Preservation and Accessibility. http://www.econtentmag.com/Articles/News/News-Feature/Archive-It-2-Internet-Archive-Strives-to-Ensure-Preservation-and-Accessibility-18132.htm. *EContent* (October 2006).

[41] Frank McCown, Sheffan Chan, Michael L. Nelson, and Johan Bollen. 2005. The Availability and Persistence of Web References in D-Lib Magazine. In *Proceedings of International Web Archiving Workshop '05*. Vienna, Austria. https://web.archive.org/web/20170924022318/http://iwaw.europarchive.org/05/papers/iwaw05-mccown1.pdf.

[42] Ian Milligan. 2012. Mining the 'Internet Graveyard': Rethinking the Historians' Toolkit. *Journal of the Canadian Historical Association* 23, 2 (March 2012), 21–64. https://doi.org/10.7202/1015788ar

[43] Ian Milligan. 2016. Lost in the Infinite Archive: The Promise and Pitfalls of Web Archives. *International Journal of Humanities and Arts Computing* 10, 1 (2016), 78–94. https://doi.org/10.3366/ijhac.2016.0161

[44] Ian Milligan. 2016. The Problem of History in the Age of Abundance. https://www.chronicle.com/article/The-Problem-of-History-in-the/238600. *The Chronicle of Higher Education* (December 2016).

[45] Tamara Munzner. 2015. *Visualization Analysis & Design*. CRC Press.

[46] David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticæ Investigationes* 30, 1 (2007), 3–26.

[47] The Library of Congress. 2017. Encoded Archival Description (Official Site). https://www.loc.gov/ead/.

[48] Abby Ohlheiser. 2017. Gothamist and DCist just abruptly shut down. What will happen to their archives? https://www.washingtonpost.com/news/the-intersect/wp/2017/11/02/gothamist-and-dcist-just-abruptly-shut-down-what-will-happen-to-their-archives/. *The Washington Post* (November 2017).

[49] Kalpesh Padia, Yasmin AlNoamany, and Michele C. Weigle. 2012. Visualizing Digital Collections at Archive-it. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '12)*. ACM, Washington, DC, USA, 15–18. https://doi.org/10.1145/2232817.2232821

[50] Peter Pirolli. 2005. Rational Analyses of Information Foraging on the Web. *Cognitive Science* 29, 3 (2005), 343–373. https://doi.org/10.1207/s15516709cog0000_20

[51] Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological Review* 106, 4 (1999), 643–675. https://doi.org/10.1037/0033-295X.106.4.643

[52] Daniel Ramage, Evan Rosen, Jason Chuang, Christopher D. Manning, and Daniel A. McFarland. 2009. Topic Modeling for the Social Sciences. In *Workshop on Applications for Topic Models, NIPS*. http://vis.stanford.edu/papers/topic-modeling-social-sciences

[53] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50. http://is.muni.cz/publication/884893/en.

[54] Rahime Belen Sağlam and Tuğba Taşkaya Temizel. 2014. Automatic information timeliness assessment of diabetes web sites by evidence based medicine. *Computer Methods and Programs in Biomedicine* 117, 2 (2014), 104 – 113. https://doi.org/10.1016/j.cmpb.2014.07.014

[55] Michael C. Sloan. 2010. Aristotle's *Nicomachean Ethics* as the Original *Locus* for the *Septem Circumstantiae*. *Classical Philology* 105, 3 (2010), 236–251. https://doi.org/10.1086/656196

[56] Storify. 2017. Storify End-of-Life. https://storify.com/faq-eol.

[57] Arlene G. Taylor. 2004. *The Organization of Information* (second ed.). Greenwood Publishing Group.

[58] Jaime Teevan, Edward Cutrell, Danyel Fisher, Steven M. Drucker, Gonzalo Ramos, Paul André, and Chang Hu. 2009. Visual snippets: summarizing web pages for search and revisitation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. Boston, MA, 20–23. https://doi.org/10.1145/1518701.1519008

[59] Edward R. Tufte. 2001. *The Visual Display of Quantitative Information* (second ed.). Graphics Press.

[60] Herber Van de Sompel, Michael L. Nelson, and Robert Sanderson. 2013. RFC 7089: HTTP Framework for Time-Based Access to Resource States – Memento. https://tools.ietf.org/html/rfc7089.

[61] Gerhard Weikum, Nikos Ntarmos, Marc Spaniol, Peter Triantafillou, András Benczúr, Scott Kirkpatrick, Philippe Rigaux, and Mark Williamson. 2011. Longitudinal Analytics on Web Archives: It's About Time!. In *Proceedings of the 5th Biennial Conference on Innovative Data Systems Research (CIDR '11)*. Asilomar, California, USA.

[62] Allison Woodruff, Andrew Faulring, Ruth Rosenholtz, Julie Morrsion, and Peter Pirolli. 2001. Using thumbnails to search the Web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '01)*. Seattle, Washington, USA, 198–205. https://doi.org/10.1145/365024.365098

[63] Renxian Zhang, Wenjie Li, and Dehong Gao. 2012. Generating Coherent Summaries with Textual Aspects. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI'12)*. AAAI Press, Toronto, Ontario, Canada, 1727–1733. http://dl.acm.org/citation.cfm?id=2900929.2900973