2018

205.3 The Many Shapes of Archive-It

Shawn Jones
Old Dominion University

Michael L. Nelson
Old Dominion University

Alexander Nwala
Old Dominion University

Michele C. Weigle
Old Dominion University

Original Publication Citation

Jones, S., Nelson, M., Nwala, A., & Weigle, M. (2018). 205.3 The many shapes of Archive-It. https://doi.org/10.17605/OSF.IO/EV42P

The Many Shapes of Archive-It

Characteristics of Archive-It Collections

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

Alexander Nwala
Old Dominion University
Norfolk, Virginia
anwala@cs.odu.edu

Michele C. Weigle
Old Dominion University
Norfolk, Virginia
mweigle@cs.odu.edu

Michael L. Nelson
Old Dominion University
Norfolk, Virginia
mln@cs.odu.edu

ABSTRACT

Web archives, a key area of digital preservation, meet the needs of journalists, social scientists, historians, and government organizations. The use cases for these groups often require that they guide the archiving process themselves, selecting their own original resources, or seeds, and creating their own web archive collections. We focus on the collections within Archive-It, a subscription service started by the Internet Archive in 2005 for the purpose of allowing organizations to create their own collections of archived web pages, or mementos. Understanding these collections could be done via their user-supplied metadata or via text analysis, but the metadata is applied inconsistently between collections and some Archive-It collections consist of hundreds of thousands of seeds, making it costly in terms of time to download each memento. Our work proposes using structural metadata as an additional way to understand these collections. We explore structural features currently existing in these collections that can unveil curation and crawling behaviors. We adapt the concept of the collection growth curve for understanding Archive-It collection curation and crawling behavior. Using the growth curves, we can see if most of the mementos in the collection are skewed earlier or later. We also introduce several seed features to describe the diversity and types of seeds present in an Archive-It collection. With these seed features, we come to an understanding of the diversity of resources that make up a collection and the depth of those resources within their seed websites, indicating whether the curator chose to preserve the top-level page or something more specific within a site. Finally, we use the descriptions of each collection to identify four semantic categories of Archive-It collections. Using the identified structural features, we reviewed the results of runs with 20 classifiers and are able to predict the semantic category of a collection using a Random Forest classifier with a weighted average F_1 score of 0.720, thus bridging the structural to the descriptive. Our method is useful because it saves the researcher time and bandwidth. They do not need to download every resource in the collection in order to identify its semantic category. Identifying collections by their semantic category allows further downstream processing to be tailored to these categories.

CCS CONCEPTS

• **Information systems → Digital libraries and archives**; *World Wide Web*;

KEYWORDS

Web Archive, Archive, Collections, Archive-It

1 INTRODUCTION

Web archiving has become an important area of digital preservation as news, research, and other content publishing has moved to the web. Government organizations seek to archive their web presence for posterity [7]. Historians [25], social scientists [8, 14], and journalists [18, 34] use web archives to understand human behavior. Because libraries have a focus on curating content specific to their communities, web archiving was once identified as a "growth area for library collections" [12]. In order to facilitate the creation of curated collections, the Internet Archive created Archive-It in 2005 [23]. Archive-It is a subscription service with the goal of allowing organizations and users to create their own web archive collections. Archive-It collections, in particular, are interesting because a single organization is responsible for each collection, meaning that the curation strategy for a collection is guided by humans rather than automated crawling operations. As web archives, Archive-It collections reflect changes to individual resources over time, providing a chronicle of unfolding world events, or the history of an organization. These changes over time make Archive-It collections different from more traditional document collections which typically contain only one version of a given resource.

How do we understand an Archive-It collection? We could look at its descriptive metadata, effectively asking others what they have said about it. We could evaluate the URI of each item in the collection to locate and then download its contents, thus **dereferencing** each URI to produce content for analysis. Such analysis can use techniques such as text mining, effectively looking at the collection's parts. We will show that other sources of information exist to provide the shapes necessary to understand the collection. What behaviors exist in web archive collections that we cannot acquire merely from metadata or text analysis? What structural features exist that can unveil curation and crawling behaviors? In this work, we examine structural features and determine what shapes exist within Archive-It collections. With each shape, we gain a better understanding of the collection.

The point of this work is to demonstrate what we can be learned through the structural features of Archive-It collections. Using only structural metadata is advantageous because it saves one from having to dereference all of the URIs in a collection in order to

understand it. Some collections have hundreds of thousands of mementos, such as the *2014 Primaries* collection with 2380 seeds and 219,084 mementos of those seeds. Our prior work summarized Archive-It collections by selecting representative items from them. These representative items were then visualized using social media storytelling [4]. That work focused mainly on collections centered on events. In this work, we identify other types of collections that may have been overlooked by our previous summarization efforts.

For a given collection, we seek to answer questions about its temporal nature. Does most of the collection exist earlier or later in its life? When did the archivist select and archive a collection's contents? Did they create a collection at start that was intended to archive new versions of the same web pages repeatedly in perpetuity? Did they nurture the selected web pages throughout the collection's life and add content continuously? Was there renewed interest at some point later in the collection's life? To answer those questions we adapt the concept of growth curves, first introduced by AlSum [5], to Archive-It collections.

We also seek to answer questions about the web pages selected for the collection. Was the collection built from web sites belonging to one organization or many? Were most of the web pages in the collection top-level pages or specific articles deeper in a web site? To answer these questions we introduce concepts like domain diversity and seed path depth diversity.

Furthermore, how well can we bridge the structural to the descriptive? We will discuss how Archive-It collections fit into four different semantic categories. As noted above, our prior work only focused on collections based on events, the smallest category. How well can we use our structural features to classify a collection into one of these semantic categories? Using the RandomForest classifier from Weka [17, 44] we show that we can predict these semantic categories with a weighted average F_1 of 0.720.

We believe that the creation of these semantic categories, as well as how well our structural features predict them, makes this work a unique contribution, because one can use these structural features to infer meaning without having to dereference all of the web pages in an Archive-It collection. This is useful because one can use this identification to support further, more specific processing tailored to that semantic category.

2 BACKGROUND

Archive-It collections consist of archived web pages, or **mementos**. An archivist creates these mementos from a list of URIs known as **seeds**, also known as "Original Resources". Thus seed selection is the genesis of a collection. The archivist then instructs a crawler [22] to create mementos from these seeds at certain intervals. The crawler produces a new version of each seed with each crawl. Depending on the chosen configuration, an archivist can also instruct the crawler to visit any additional web pages linked to from the seeds. This process produces even more mementos. To reduce confusion in this work, we will use the term **seed memento** to specifically refer to mementos created from seeds. Seed mementos are of particular interest because they tie back to decisions made explicitly by the archivist and thus represent unique policy and behavior for each collection.

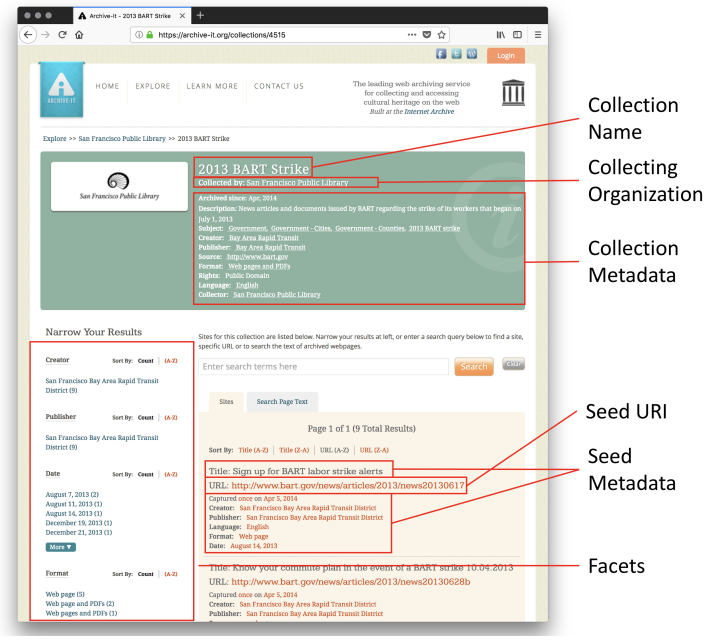


Figure 1: Example Collection from https://archive-it.org/collections/4515

Archive-It's user interface for each collection can provide the viewer a wealth of information, as shown in Figure 1, including a description, collection metadata, seed metadata, and facets for discovery and searching. The archivist is only required to supply the collection name. The collecting organization in the *collected by* field is drawn from the archivist's account. Collections can be marked public, in which seed URIs and seed metadata is available, or private, where only collection metadata is available. Each collection includes an *archived since* field, indicating when the collection was created.

With the exception of the collection name, the collecting organization, and the *archived since* field, all additional metadata is optional and provided by the archivist. For each collection, the archivist can choose from a controlled vocabulary of fields based on Dublin Core [35] or they can use their own freeform vocabulary. These same fields can be applied individually for each seed. Archivists may also select one or more topics for the collection. Like the metadata fields, there is a controlled vocabulary, but archivists may add additional topics as well. Unfortunately, this metadata is inconsistently applied to collections, likely due to differences in content standards and rules of interpretation among archivists.

Being compliant with the Memento Protocol [42], each Archive-It collection also provides a **TimeMap** for each seed. TimeMaps are lists of mementos for a specific seed. Using TimeMaps, one can acquire the URIs of all mementos for a given seed as well as each memento's **memento-datetime**, the datetime that the memento was recorded.

3 RELATED WORK

Digital collections, and more specifically, those at Archive-It, have been explored by others in the past.

Fenlon explores the different types of digital collections that exist [16]. She mentions that a digital collection's contents, metadata, and even user interface are constructed based on the needs of the audience that they serve. Even though she did not focus on web archives, her work is related because it indicates that there is additional appetite among scholars to understand the features of digital collections outside of our summarization efforts.

Milligan [26] reviews three web archive collections to determine the effectiveness of different techniques for choosing seeds. The three collections differed in how seeds were chosen: (1) through seed URIs extracted from tweets within a given Twitter hashtag, (2) from general crawling via the Internet Archive, or (3) manually by curators at the University of Toronto Libraries. He discovered that a combination of hashtags and careful curation proves best. It is the behavior of those who create this third type of collection that we study in this paper. Likewise, Nwala evaluates how to use search engine result pages (SERPs) to discover news stories appropriate for building or augmenting web archive collections [31]. Our work differs because we analyze the seeds after selection.

Work has been done to understand the behaviors of the users who create collections with live web curation platforms. Using the Uses and Gratifications model [40], Mull [27] discovered the following motivations for using the image curation platform Pinterest: "fashion", "creative projects", "entertainment", "virtual exploration", and "organization". Wang [43] applied the MAIN model [41] to explain the different gratifications of Pinterest users in an attempt to understand why users engaged with the platform. The results of Wang's study indicate that Pinterest users create collections for the purpose of engaging with the topics that they find to be fun and exciting, in pursuit of escapism. The users analyzed in these studies are different from the institutions that create collections in Archive-It. Those institutions have different motivations for creating a collection. Some have legal requirements to archive government agencies. Others collect resources on behalf of scholars at the institution. Our work involves understanding what behavior can be derived from the features in web archive collections rather than conducting user studies to understand their motivations for creating collections.

Ogden brings to light the behavior and work of web archivists [33]. She applies an ethnographic approach to understand those who participate in the work of the Internet Archive, noting that they are currently focused on methods for discovering seeds. Crook used Archive-It as part of an effort to produce a web archive of online Australian publications [13]. He highlights the challenges of using the Archive-It service, especially for those used to having complete control over the archiving and playback processes. Slania [39] and Deutch [15] detail the challenges of using Archive-It to archive art web sites. Where their work focuses on the impressions of web archivists, ours focuses on studying the output of their work. We review their behavior as observed from the structural features of Archive-It collections themselves.

Niu evaluated the capabilities of ten different web archives, including Archive-It, highlighing features such as keyword searching and date facets [29]. Rather than criticizing or evaluating the capabilities of Archive-It, our work is intended to highlight structural features that may be used to better understand its collections.

Encoded Archival Description (EAD) [32] is an "XML standard for encoding archival finding aids" maintained by the Library of Congress. Archive-It does not currently use EAD, instead favoring a metadata scheme based on Dublin Core [6]. Our work attempts to identify structural features that exist within web archive collections rather than relying upon existing metadata.

AlNoamany evaluated different methods of detecting off-topic pages within web archives, focusing on Archive-It [3]. 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 [38]. We are looking at structural features rather than the content of the web archive collections themselves.

Abramson focuses on machine learning techniques that can classify URIs based on their contents without dereferencing (downloading) them [1]. Though it could be used to augment our summarization work, we currently focus on other aspects of URIs and their diversity within a collection.

AlSum analyzes different web archives to determine which seeds they cover over which periods of time [5]. To acquire seeds, the authors randomly sampled URIs from DMOZ as well as search engines provided by the web archives themselves. AlSum discusses the age of the mementos within each archive, which top-level domains are covered, which languages exist within the mementos of the archive, and the growth curves of each corpus over time. We use these growth curves in our own work. We also review all seeds and seed mementos available within Archive-It, focusing on the growth curves at a collection level rather than at the level of an entire web archive.

Like this work, AlNoamany also reviews some characteristics of archived collections within Archive-It [2]. That analysis, performed in 2015, focused on the number of seeds, the mean number of mementos per seed, the time span of each collection, the domains within these collections, which top level domains were represented, and the decay rate of resources within all Archive-It collections. Our work is different because we look at collections as units unto themselves and have developed different metrics to measure them.

4 DATA ACQUISITION

Our data breakdown for this work is shown in Table 1. To acquire seed URIs, we used BeautifulSoup [28, 36] to scrape the web pages of 9,351 Archive-It collections between October and December of 2017. Because private collections do not expose seeds, we were only able to acquire the public Archive-It collections. For each public collection there also exists a comma-separated report of seed URIs. In 6% of collections, the number of seeds acquired between web scraping and this report did not match. To account for these cases, we used the union of these two sources to get the list of seeds for each collection.

We did not use the Archive-It CDX/C API because it requires knowing seed URIs beforehand [9] and we did not use the optional OAI-PMH interface available for many collections [10] because it would not provide all of the information we needed. Also, not all collections have enabled this feature.

Table 1: Reduction of Data for this Paper

Data Category Description	Count	Remaining
Total Collections	9,351	
Removed Private Collections	4,823	4,528
Removed Collections Archived Since Jan, 2017	440	4,088
Removed Collections With No Mementos	248	3,840
Removed Collections With Errors	48	3,792
Removed Singletons	357	3,435
Removed Single Second Collections	21	3,414
Removed Test Collections	32	3,382
Collections Remaining For Analysis		3,382

Once we had the seeds, we extracted domain names using tldextract [19]. We acquired the seed memento URIs and their memento-datetimes via Memento TimeMaps. Because they are effectively empty, we removed collections from our dataset that had no mementos. Some of these TimeMaps were empty due to downloading or processing errors, and thus their collections were removed. We also removed all collections containing **singletons**, consisting of a single seed with a single memento, because they do not provide enough information to create growth curves. Likewise, we removed collections where all mementos were crawled within the same second. For semantic analysis, we reviewed each collection's name and associated metadata and removed collections that were marked by the archivist with the terms *test* or *trial*. This left us with 3,382 collections consisting of 700,835 seeds and 6,943,677 seed mementos to review.

5 STRUCTURAL FEATURES

5.1 Collection Growth Curves

A collection growth curve provides insight into the seed curation and crawling behavior of an Archive-It collection. Figure 2 shows an example growth curve for Archive-It Collection 366[1]. The x-axis represents the **life of the collection**, or the time between a collection's first memento and its last. To normalize among collections with different durations, we show the x-axis as a percentage of the collection's lifespan. The y-axis represents the percentage of URIs in the collection at a given time. URIs come in two categories: seeds or seed mementos, represented by the green and red lines, respectively.

Growth curves for Archive-It collections consist of multiple parts. Figure 3 demonstrates how to interpret the information within a growth curve. An imaginary diagonal line shows a linear relationship between the growth of URIs over time. It divides each graph into two parts. If the seed line (green) is in the upper left corner, then most of the seeds were added earlier to the collection, and if

[1]Archive-It collections can be accessed by appending the collection number to the URI https://archive-it.org/collections/, so collection 366 would be https://archive-it.org/collections/366

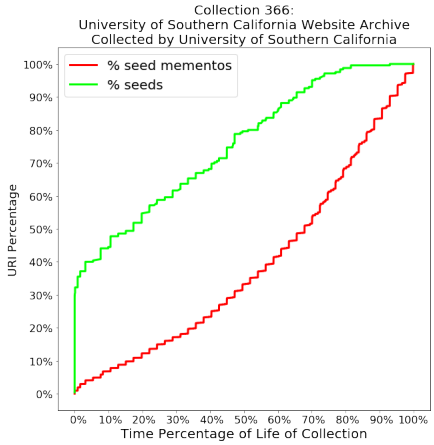


Figure 2: The Growth Curve of Archive-It Collection 366

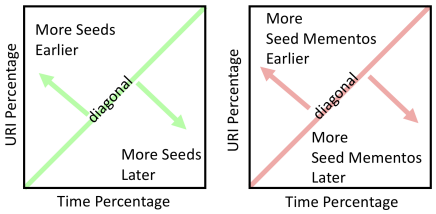


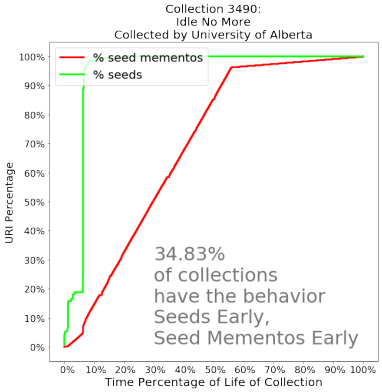
Figure 3: The Anatomy of a Collection Growth Curve

the seed line occupies the lower right corner, then most were added later. The seed line reflects an aspect of curatorial engagement with the collection, indicating when the archivist first crawled a given seed. The closer the seed line is to the diagonal, the more often the archivist added a new seed. The memento-datetime of the first memento for each seed is used to generate the seed line.

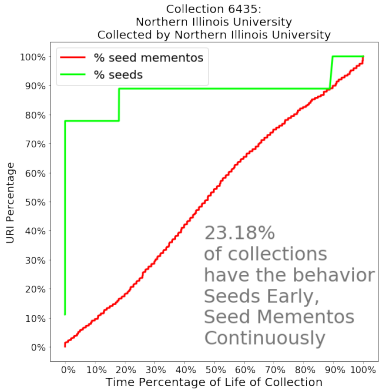
The meaning of the seed memento line is similar. Where the seed line indicates intent, the seed memento line indicates the growth of actual collection. If the seed memento line (red) mostly occupies the upper left corner, then most of the mementos were crawled earlier in the collection's life, meaning that most of the collection's holdings were created at that time, and its temporality is skewed earlier. If the seed memento line occupies the lower right corner, then the collection's temporality is skewed later. If the seed memento line runs along the diagonal, then the collection's temporality is spread more evenly across the collection. The memento-datetimes of all mementos are used to generate the seed memento line.

5.1.1 Interesting Growth Curve Behaviors. Using the area under the curve (AUC), we were able to identify different collection behaviors. If the AUC exceeds 0.5 — the area of the diagonal — then the growth was earlier. If the AUC is less than 0.5 then the growth was later. If the AUC is within 0.05 of the diagonal, then we considered it growing *continuously*. Using these three primitives, we identified the behaviors shown in Figure 4.

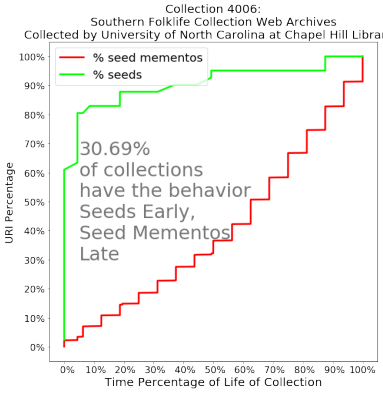
Seeds early, seed mementos early — Seen in Figure 4a with collection *Idle No More* (ID 3490), the growth curves with this behavior indicate that the archivist made most curatorial decisions



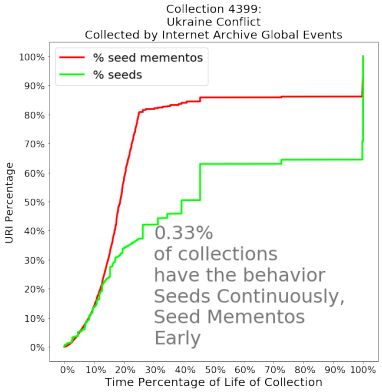
(a)
Archive-It Collection 3490:
Seeds early,
seed mementos early



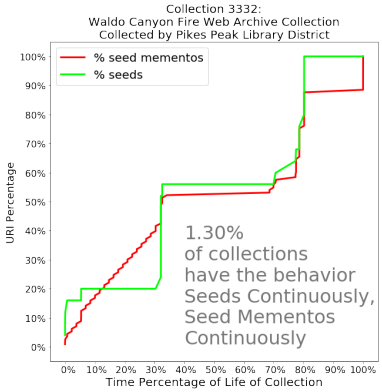
(b)
Archive-It Collection 6435:
Seeds early,
seed mementos continuously



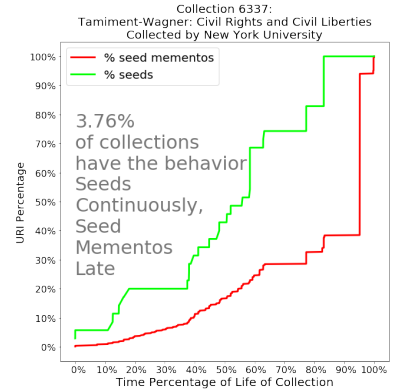
(c)
Archive-It Collection 4006:
Seeds early,
seed mementos late



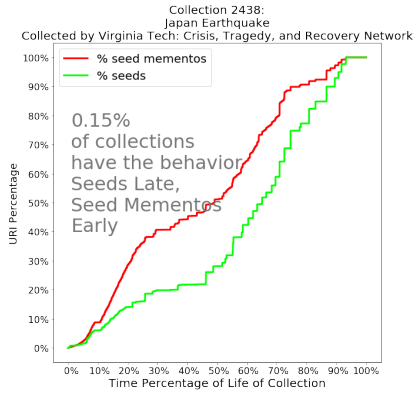
(d)
Archive-It Collection 4399:
Seeds continuously,
seed mementos early



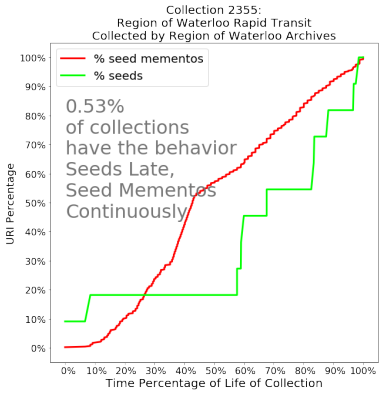
(e)
Archive-It Collection 3332:
Seeds continuously,
seed mementos continuously



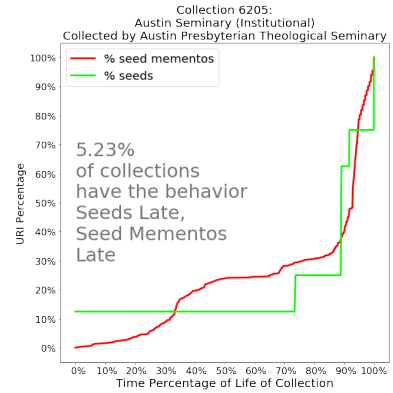
(f)
Archive-It Collection 6337:
Seeds continuously,
seed mementos later



(g)
Archive-It Collection 2438:
Seeds Late,
seed mementos early



(h)
Archive-It Collection 2355:
Seeds late,
seed mementos continuously



(i)
Archive-It Collection 6205:
Seeds late,
seed mementos late

Figure 4: Examples from nine different growth curve behavior categories, grey inset text conveys the percentage of the 3,382 collections in this study

near the start of the life of the collection. The seed memento line is skewed early, indicating that most of the seed mementos in the collection come from that time period.

Seeds early, seed mementos continuously — Figure 4b shows collection *Northern Illinois University* (ID 6435), where the archivist added more than 70% of the seeds near the beginning of the collection's life. The archivist selected seeds early, but then chose crawling strategies that added seed mementos steadily throughout its existence.

Seeds early, seed mementos late — In all of these cases, the archivist chose seeds early, but the crawling strategy produced seed mementos at a later time. In collection *Southern Folklife Collection Web Archives* (ID 4006), shown in Figure 4c, we see a case where the archivist added 60% of the seeds earlier in the collection's life. The crawling strategy ensured that seed mementos were crawled later. In this case 50% of all mementos were crawled by 65% of its life.

Seeds early is the most frequent seed behavior, taking place in 88.7% of all collections studied.

Seeds continuously, seed mementos early — Figure 4d shows collection *Ukraine Conflict* (ID 4399), where the seed growth curve wraps around the diagonal, indicating that the archivist added seeds more regularly, but most crawling happened earlier in the collection's life. This means that there are more seed mementos from the earlier seeds.

Seeds continuoulsy, seed mementos continuously — Collection *Waldo Canyon Fire Web Archive Collection* (ID 3332) is shown in Figure 4e. Collections with this behavior indicate continuous involvement both on the part of seed selection as well as crawling. Both lines wrap the diagonal as the collection grows steadily.

Seeds continuously, seed mementos late — Shown in Figure 4f with collection *Tamiment-Wagner: Civil Rights and Civil Liberties* (ID 6337), this behavior indicates that the archivist was continuously engaged in adding seeds to the collection, but most of the mementos were created later in the collection's life.

Seeds late, seed mementos early — This behavior is demonstrated by collection *Japan Earthquake* (ID 2438) in Figure 4g. In this case, the seed memento growth line exists farther left on the graph than the seed growth line. The early seeds added to the collection have more memento growth than the seeds that follow, because the archivist added more seeds later in the collection's life.

Seeds late, seed mementos continuously — Figure 4h shows collection *Region of Waterloo Rapid Transit* (ID 2355). In this case, the seed memento growth is steady, but something changed around 60% of its life span. Approximately 60% of the seed mementos belong to the first 20% of seeds.

Seeds late, seed mementos late — This behavior is exemplified by collection *Austin Seminary (Institutional)* (ID 6205), shown in Figure 4i. In this case, more seeds were added when the collection was already 70% old. Another dramatic shift happened when more seeds were added at the 90% mark and again later. This could indicate dramatically renewed interest in this collection.

Memento growth overtakes seed growth — Sometimes the seed growth stops for a while and the seed memento growth overtakes it, visualized as the seed memento line being higher than the seed line. Figures 4d, 4g, 4h, and 4i all demonstrate this behavior.

All seeds up front — All of the seeds in these collections were chosen at the beginning of the collection's life, but the seed memento growth varies. In these cases, the archivist not only chose the seeds early in the collection's life, but they never added seeds later. It is an extreme case of the seed line from Figure 4a.

Curve overlapping — Sometimes the ratio of seeds to seed mementos is 1-to-1 or very close to 1-to-1. The lines in these cases show up as overlapping. Their growth behavior indicates that very few crawls are happening per seed. In these cases, it is likely that the archivist just wanted to gather a single copy of a given seed rather than recording the changes to a seed over time. Because we removed singletons from the dataset, all of these collections have more than one seed.

5.1.2 Growth Curve Features. We have identified five growth curve features which provide insight into the behavior of a collection.

The **number of seeds** submitted to the collection varies, as does the **number of seed mementos**. These can be counted by using the seed acquisition activities and TimeMaps mentioned above.

Difference between seed curve AUC and diagonal — The AUC of the seed curve indicates whether the seeds were added earlier or later to the collection. Subtracting this value from the AUC of the diagonal gives additional information useful for understanding the nature of the seed curve. Negative values indicate that seeds were added later. Positive values indicate that seeds were added earlier. Values very close to 0 indicate that seeds were added continuously.

Difference between seed memento curve AUC and diagonal – The area under the seed memento curve is useful as well. Subtracting this value from the AUC of the diagonal provides similar information to the seed AUC feature mentioned above.

Difference between seed curve AUC and seed memento curve AUC — Subtracting the AUC of both curves indicates how close they are to each other. A value of 0 indicates that the curves are overlapping, likely meaning that there is one memento per seed. A positive value means that the seeds are added earlier than the seed mementos. A negative value means that the seed memento growth has overtaken the seed growth.

Collection Lifespan — The collection lifespan is the difference in memento-datetime between the last memento and the first.

5.2 Seed Features

In addition to the growth curves, structural features of the seeds themselves provide insight into the behavior of the archivist with respect to a collection.

Seed URI domain diversity — Seed URI domain diversity [30] 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. Equation 1 computes the diversity D as the number of unique domains U divided by the number of seeds C. In Equation 2 D' normalizes this diversity value D between 0 and 1. A collection with 1 seed, by definition, has a diversity of 0.

Table 2: 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%

$$D = \frac{U}{C} \tag{1}$$

$$D' = \frac{CD - 1}{C - 1} = \frac{U - 1}{C - 1} \tag{2}$$

The **path depth** for each seed URI consists of the number of items separated by slashes after the domain name. If the path consists of a query string, a 1 is added to the path depth, similar to [24]. If the last item in the path consists of a known default page (e.g., index.html), then we subtract 1 from the path depth. Default pages are determined by a list of well known default pages[2]. Path depths indicate crawling intention by the archivist with respect to the collection. Recall that seeds are the starting point for a crawl, thus an archivist who selects a path depth of 0 seeks to start crawling from the top of a web site, whereas one who starts with a higher path depth may be starting with a page containing more specific content.

Seed URI path depth diversity — We acquire an idea of the spread of path depth across the collection by applying the above domain diversity equation to the seed path depth of every seed in the collection. This may indicate if the seed URIs consist solely of top-level pages or a mixture of top-level pages and more specific content.

Most frequent seed URI path depth — If a collection's most frequent seed URI path depth is 0, then it mostly consists of seeds of web site top-level pages. If the most frequent path depth is higher, then it mostly consists of seeds deeper in a web site.

% Query string usage — Some collections consist mostly of URIs with query strings, whereas others consist of just paths. A collection with 1 seed has either 0% or 100% query string usage.

6 BRIDGING THE STRUCTURAL TO THE DESCRIPTIVE

Each structural feature tells part of the story of a collection. We also wanted to know how well these features mapped to the descriptive metadata for each collection. Our goal is to be able to predict some aspect of the descriptive information from the structural features introduced in the last section.

Every Archive-It collection has one or more assigned topics. Some of these topics come from a controlled vocabulary, but the archivist has the option of providing one or more freeform topics of their own as well. We evaluated several classifiers to predict these controlled vocabulary topics. The best weighted average F_1 score we achieved was 0.225 with the Logistic Model Tree classifier [20]. The poor results were likely due to the one-to-many relationship of these archivist-assigned topics.

Considering that the user-supplied topics were not suitable for prediction, we reviewed the Archive-It collections by hand and placed them into four **semantic categories**. The distribution of these semantic categories is shown in Table 2.

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* (ID 2278) archived by the University of Utah, or the *City of Eagan Websites* (ID 2289) archived by the City of Eagan, Minnesota. In each case the organization is archiving its own web presence. Less apparent are collections like *Governor of Tennessee, Phil Bredesen* (ID 391) archived by the Tennessee State Library and Archives. In these latter cases, the archiving organization, the name of the collection, and/or the ownership of the seeds do not match, but the Tennessee State Library and Archives specifically exists to archive the State of Tennessee government. Tennessee State Libraries has collections for many, if not all, Tennessee state agencies. From this behavior, we can infer that they are tasked with archiving the web presence of all Tennessee state government. Other organizations with collections that fit into this category are the Federal Depository Library Program Web Archive and the Region of Waterloo Archives.

Subject-based Archiving — Some collections consist of a number of seeds bound by a single topic. The topic may be clear, as with *Environmental Justice* (ID 7635), archived by Tufts University. The topic may also be vague, like *ISU Special Collections Department Manuscript Collections Web Sites* (ID 1501) archived by Iowa State University (ISU). In the former, the subject is in the collection name. In the latter the subject is organizations that have shared physical items with the ISU Special Collections Department. This is not Self-Archiving because these organizations are not part of ISU, nor is it apparent that ISU is specifically tasked via some broader archiving initiative to archive them. ISU is merely complementing their physical library collection by archiving additional information about the organizations who have contributed to it.

Time Bounded - Expected — These collections focus on an expected, planned event, such as *2008 Olympics* (ID 871) 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* (ID 663) archived by the Library of Virginia. Collections from institutions participating in the K-12 Archiving Initiative [21] also fit into this category, as they are planned to exist for a single semester or school year.

Time Bounded - Spontaneous — These collections start after a spontaneous event. Collections fitting into this category include *Tucson Shootings* (ID 2305) archived by the Virginia Tech: Crisis, Tragedy, and Recovery Network and *2011 Japan Earthquake* (ID 2450) archived by the University of Michigan, School of Information. They may also start after the beginning of a movement, such as *Black Lives Matter Movement* (ID 6396) archived by the San Jose State University, School of Information. The key is that these collections are started due to this spontaneous event or movement and are usually terminated at some point.

[2]https://support.tigertech.net/index-file-names

Table 3: Weighted average results of 10 best classifiers for predicting semantic class evaluated using 10-fold cross validation runs while training on the complete data set

	Weighted Average			
Classifier	**TPR**	**FPR**	F_1	**ROC Area**
Random Forest	0.728	0.182	0.720	0.871
ForestPA	0.713	0.201	0.701	0.854
Decision Table	0.702	0.214	0.685	0.831
LMT	0.702	0.205	0.685	0.833
CDT	0.700	0.212	0.686	0.819
JRip	0.698	0.235	0.679	0.769
Simple Cart	0.694	0.199	0.683	0.811
FT	0.693	0.201	0.681	0.789
BFTree	0.689	0.214	0.676	0.766
Multilayer Perceptron	0.686	0.197	0.675	0.818

We determined that Random Forest [11] is the best classifier for predicting the semantic category. We arrived at this conclusion by testing 20 classifiers with Weka v. 3.8.2 [17]. We set the semantic category as the target class and evaluated several machine learning algorithms with 10-fold cross validation. Table 3 shows the weighted average results of classification runs using these structural features for the top 10 classifiers we evaluated. We have four classes, so these weighted average results do not provide a complete picture. Table 4 shows the results for Random Forest by semantic category. Self-archiving scores the best, with an weighted average F_1 score of 0.847. This is likely due to the fact that 54.1% of Archive-It collections fall into this category. Other semantic categories do not fare so well. Time Bounded - Spontaneous is the worst, with F_1 of 0.456. This is likely because it only makes up 4.2% of all collections, giving the classifier little with which to train. More surprising is that Time Bounded - Expected does so well at 0.621, even though it only makes up 14.1% of all collections. Finally, Subject-based Archiving is slightly worse with an F_1 of 0.562 in spite of making up 27.6% of all collections. Random Forest also had the best F_1 score of all classifiers in all semantic categories. Thus, Random Forest trained on our data set performs best at identifying to what semantic category a collection belongs.

Better results for almost all semantic categories can be achieved by removing the number of mementos feature. We converted each of the four semantic categories to a numeric value of 1 - 4 and then calculated Kendall's τ on the feature-category combination, with the results shown in Table 5. Domain diversity and collection lifespan have the highest correlation to the categories, with scores of 0.3863 and -0.3320, respectively. The number of mementos and the most frequent URI depth have lowest correlation. We removed the lowest-scoring attributes one at a time. Removing the number of mementos feature and retraining with Random Forest, shown in Table 6, improved the F_1 scores of Subject-based Archiving, Time Bounded - Expected, and Time Bounded - Spontaneous to 0.568, 0.641, and 0.462 respectively. The score of Self-Archiving went down from 0.847 to 0.841. Removing more features and retraining did not improve the F_1 scores further.

Table 4: Results by class of Random Forest classifier for predicting semantic category evaluated using 10-fold cross validation runs

	Weighted Average			
Semantic Category	**TPR**	**FPR**	F_1	**ROC Area**
Self-Archiving	0.891	0.250	0.847	0.899
Subject-based Archiving	0.538	0.144	0.562	0.794
Time Bounded - Expected	0.588	0.050	0.621	0.911
Time Bounded - Spontaneous	0.364	0.010	0.456	0.879
Weighted Average	0.728	0.182	0.720	0.871

Table 5: Kendall τ Correlation of Features to Semantic Categories

Feature	**Kendall τ**
Seed URI domain diversity	0.3863
Collection lifespan	-0.3320
Number of seeds	0.2878
Difference between seed curve AUC and seed memento curve AUC	-0.2416
% query string usage	0.2265
Difference between seed memento Curve AUC and diagonal	0.1569
Difference between seed curve AUC and diagonal	-0.1387
Seed URI path depth diversity	0.1317
Most frequent seed URI path depth	-0.0687
Number of mementos	0.0561

Table 6: Results by class for Random Forest classifier with the number of mementos feature removed

	Weighted Average			
Semantic Category	**TPR**	**FPR**	F_1	**ROC Area**
Self-Archiving	0.881	0.251	0.841	0.891
Subject-based Archiving	0.549	0.146	0.568	0.782
Time Bounded - Expected	0.609	0.048	0.641	0.906
Time Bounded - Spontaneous	0.364	0.009	0.462	0.877
Weighed Average	0.729	0.183	0.722	0.862

7 FUTURE WORK

As noted in the introduction, we intend to adapt these structural features and our classifier results to better support our collection summarization work [4]. For example, if growth curves indicate that a collection's mementos are skewed earlier, we can select different mementos for our storytelling summarization. The seed analysis features provide information on the diversity of a collection, allowing us to change our algorithms to better choose which seeds are included. Using the classifier, we can tailor summarization algorithms to specific semantic categories of Archive-It collections. We can also augment these features with semantic information as well, such as by analyzing the seed URIs with Abramson's method [1].

Although we were able to extract most of our needed data using Memento and screen scraping, a structured metadata API would

have been helpful. We intend to work further with Archive-It to develop this capability via the WASAPI project [37] or similar work.

8 CONCLUSIONS

Archive-it collections can be understood, but not only via their metadata or contents. We have shown that there exist structural features that provide additional information on the shapes necessary to understand a collection. In addition to the number of seeds and number of seed mementos, more complex features exist.

We have adapted collection growth curves to Archive-It collections, revealing their behaviors. Collection growth curves consist of two lines, a seed line and a seed memento line. The seed line indicates when a seed was first added to a collection. The seed memento line conveys the crawling behavior for all seeds in a collection over time. Using these two curves, we can see if the archivist controlling a collection added seeds early, late, or continuously, indicating the level of curatorial involvement with the collection. Likewise, we can see if the seed mementos were crawled early, late, or continuously, indicating the crawling strategy of a collection. We discovered that most collections have their seeds skewed early. Through these curves we gain an understanding of the skew of the temporality of a collection.

We have also identified seed features that help with understanding the curation strategy of a collection. Via domain diversity, we can tell if a collection consists of seeds from one domain or many, thus understanding that the collection comes from many different sources. Using the most frequent URI path depth, we determine if most of the collection consists of top-level pages or specific deep URIs. With seed path depth diversity, we understand the spread of path depths within a collection, indicating if most of its seeds have the same path depth. Understanding how much of a collection uses a query string in its URIs also provides information on the nature of its seeds. Through these features, we gain an understanding of the nature of what was chosen for archiving.

We bridged the structural and the descriptive by classifying collections into four semantic categories: Self-Archiving, Subject-based Archiving, Time Bounded - Expected, and Time Bounded - Spontaneous. We discovered that Self-Archiving is the most prevalent semantic category, making up 54.1% of the collections surveyed. We also provided the results of training runs with classifiers, and determined that the Random Forest classifier performs best at identifying the semantic category, with a weighted average F_1 score of 0.720. We discussed how one could improve the scores of the classifier for most semantic categories by removing the number of mementos feature.

In this work, we have shown that semantic metadata and the collection's holdings themselves are not the whole picture and that there are many shapes to Archive-It.

ACKNOWLEDGMENTS

This work supported in part by the Institute of Museum and Library Services (LG-71-15-0077-15).

REFERENCES

[1] Myriam Abramson and David Aha. 2012. What's in a URL? Genre Classification from URLs. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence, Palo Alto, California.

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

[3] Yasmin AlNoamany, Michele C. Weigle, and Michael L. Nelson. 2016. Detecting off-topic pages within TimeMaps in Web archives. *International Journal on Digital Libraries* 17, 3 (2016), 203–221. https://doi.org/10.1007/s00799-016-0183-5

[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, 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

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

[7] The National Archives. 2018. UK Government Web Archive - The National Archives. http://www.nationalarchives.gov.uk/webarchive/.

[8] William Y. Arms, Selcuk Aya, Pavel Dmitriev, Blazej Kot, Ruth Mitchell, and Lucia Walle. 2006. A Research Library Based on the Historical Collections of the Internet Archive. http://www.dlib.org/dlib/february06/arms/02arms.html. *D-Lib Magazine* 12, 2 (February 2006).

[9] Karl-Rainer Blumenthal. 2017. Access Archive-It's Wayback index with the CDX/C API. https://support.archive-it.org/hc/en-us/articles/115001790023-Access-Archive-It-s-Wayback-index-with-the-CDX-C-API.

[10] Karl-Rainer Blumenthal. 2017. Access web archives with the OAI-PMH metadata feed. https://support.archive-it.org/hc/en-us/articles/210510506-Access-web-archives-with-the-OAI-PMH-metadata-feed.

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

[12] Daniel Chudnov. 2011. Saving the Web. https://www.questia.com/magazine/1P3-2538290041/saving-the-web. *Computers in Libraries* 31, 10 (December 2011), 30 – 32.

[13] Edgar Crook. 2009. Web archiving in a Web 2.0 world. *The Electronic Library* 27, 5 (2009), 831–836. https://doi.org/10.1108/02640470910998542

[14] 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

[15] Samantha Deutch and Sally McKay. 2016. The Future of Artist Files: Here Today, Gone Tomorrow. *Art Documentation: Journal of the Art Libraries Society of North America* 35, 1 (2016), 27–42. https://doi.org/10.1086/685975

[16] Katrina Fenlon. 2017. Toward a characterization of digital humanities research collections: A contrastive analysis of technical designs. *Proceedings of the Association for Information Science and Technology* 54, 1 (2017), 82–92. https://doi.org/10.1002/pra2.2017.14505401010

[17] 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.

[18] 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).

[19] John Kurkowski. 2017. tldextract. https://github.com/john-kurkowski/tldextract.

[20] Niels Landwehr, Mark Hall, and Eibe Frank. 2005. Logistic Model Trees. *Machine Learning* 59, 1 (01 May 2005), 161–205. https://doi.org/10.1007/s10994-005-0466-3

[21] Cheryl Lederle. 2016. Your Students Can Archive the Internet — Apply Now. https://blogs.loc.gov/teachers/2016/07/your-students-can-archive-the-internet-apply-now/.

[22] Jillian Lohndorf. 2017. Archive-It Crawling Technologys. https://support.archive-it.org/hc/en-us/articles/115001081186-Archive-It-Crawling-Technology.

[23] 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).

[24] 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 IWAW'05*. Vienna, Austria. http://iwaw.europarchive.org/05/papers/iwaw05-mccown1.pdf.

[25] 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).

[26] Ian Milligan, Nick Ruest, and Jimmy Lin. 2016. Content Selection and Curation for Web Archiving: The Gatekeepers vs. The Masses. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries (JCDL '16)*. ACM, Newark, New Jersey, USA, 107–110. https://doi.org/10.1145/2910896.2910913

[27] Ian R. Mull and Seung-Eun Lee. 2014. "PIN" pointing the motivational dimensions behind Pinterest. *Computers in Human Behavior* 33 (2014), 192 – 200. https:

//doi.org/10.1016/j.chb.2014.01.011
[28] Vineeth G. Nair. 2014. *Getting Started with Beautiful Soup.* Packt Publishing.
[29] Jinfang Niu. 2012. Functionalities of Web Archives. https://doi.org/10.1045/march2012-niu2. *D-Lib* 18, 3/4 (March/April 2012).
[30] Alexander Nwala. 2018. 2018-05-04: An exploration of URL diversity measures. https://ws-dl.blogspot.com/2018/05/2018-05-04-exploration-of-url-diversity.html.
[31] Alexander C. Nwala, Michele C. Weigle, and Michael L. Nelson. 2018. Scraping SERPs for Archival Seeds: It Matters Where You Start. In *Proceedings of the 18th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '18)*. ACM, June, 263–272.
[32] The Library of Congress. 2017. Encoded Archival Description (Official Site). https://www.loc.gov/ead/.
[33] Jessica Ogden, Susan Halford, and Leslie Carr. 2017. Observing Web Archives: The Case for an Ethnographic Study of Web Archiving. In *Proceedings of the 2017 ACM on Web Science Conference (WebSci '17)*. ACM, New York, NY, USA, 299–308. https://doi.org/10.1145/3091478.3091506
[34] 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).
[35] Maria Praetzellis. 2017. Archive-It: Add, edit, and manage your metadata. https://support.archive-it.org/hc/en-us/articles/208332603-Add-edit-and-manage-your-metadata.
[36] Leonard Richardson. 2017. Beautiful Soup Documentation. https://www.crummy.com/software/BeautifulSoup/bs4/doc/.
[37] David S. H. Rosenthal, Jefferson Bailey, and Nicholas Taylor. 2017. Building API-Based Web Archiving Systems and Services. https://archive.org/details/BaileyWASAPIAPIs2017.
[38] 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
[39] Heather Slania. 2013. Online Art Ephemera: Web Archiving at the National Museum of Women in the Arts. *Art Documentation: Journal of the Art Libraries Society of North America* 32, 1 (2013), 112–126. https://doi.org/10.1086/669993
[40] Thomas F. Stafford, Marla Royne Stafford, and Lawrence L. Schkade. 2004. Determining Uses and Gratifications for the Internet. *Decision Sciences* 35, 2 (2004), 259–288. https://doi.org/10.1111/j.00117315.2004.02524.x
[41] S Shyam Sundar. 2008. The MAIN model: A heuristic approach to understanding technology effects on credibility. *Digital media, youth, and credibility* 73–100 (2008).
[42] 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.
[43] Ruoxu Wang, Fan Yang, Saijing Zheng, and S. Shyam Sundar. 2016. Why Do We Pin? New Gratifications Explain Unique Activities in Pinterest. *Social Media + Society* 2, 3 (2016), 2056305116662173. https://doi.org/10.1177/2056305116662173
[44] Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal. 2016. *Data Mining: Practical Machine Learning Tools and Techniques* (fourth ed.). Morgan Kaufmann. ISBN: 978-0128042915.