Journal of Contemporary Archival Studies

Volume 7 Article 19

2020

Labor Gone Digital (DigiFacket)! Experiences from Creating a Web Archive for Swedish Trade Unions

Jenny Jansson
Uppsala University, jenny.jansson@statsvet.uu.se

Katrin Uba
Uppsala University, katrin.uba@statsvet.uu.se

Jaanus Karo
Uppsala University, karo.jaanus@gmail.com

Labor Gone Digital (DigiFacket)! Experiences from Creating a Web Archive for Swedish Trade Unions

Cover Page Footnote

This project is funded by the Swedish Foundation for Humanities and Social Sciences (Riksbankens jubileumsfond, IN14-0698:1).

Introduction

The Internet has become an increasingly important forum for societal activism, as event mobilization, member organization, and some actions (e.g., petitioning) have moved online.[1] These new types of activities, often facilitated by diverse social media platforms (such as Facebook, Twitter, and Instagram) form an increasingly important part of organizations' work and communication. These texts, photos, audio recordings, and videos should be preserved for posterity. However, even though web archiving has rapidly increased in recent years, much of the born-digital material that is generated and available on the Internet is still not systematically archived.[2] This is especially true for civil society groups such as labor organizations. There are several reasons for this: scarce resources, the lack of common standards for born-digital record management, and the failure to appreciate the material's value by the organizations themselves.[3] Some of this material may be considered as mere communication by the movements and thus may not be seen as worth preserving. Unlike government agencies and authorities, where legislation regulates the preservation of materials, there are no corresponding requirements for civil society organizations.

In countries like Sweden, which has a long tradition of archiving materials from social movements and other civil society organizations, this development has become a significant concern for researchers. Born-digital material is very perishable and is often only temporarily available online.[4] Moreover, some material is held in the depositories of private enterprises (e.g., Facebook and Twitter), making it difficult to foresee what the access options will be in the future. If nothing is done soon, it may be very difficult to investigate our contemporary organizations and social movements in the future.[5]

The only remedy to this situation is the development of various forms of e-archives. Indeed, there have been several web archiving initiatives to preserve websites and other born-digital material, the most famous one being the Internet Archive.[6] By necessity, these important but large web-archiving projects are rough in their configurations: for example, there may be irregular and sparse downloads, limitations to certain top-level domains, and limited search options. For instance, the National Library of Sweden's project Kulturarw3 downloads Swedish websites twice a year, which provides a good overview of Swedish websites, but the infrequent downloads create troublesome gaps for researchers.[7] Moreover, although some scholars have proposed ways to counter the problem of archiving the materials from social

[1] Victoria Carty and Francisco G. R. Barron, "Social Movements and New Technology: The Dynamics of Cyber Activism in the Digital Age," in *The Palgrave Handbook of Social Movements, Revolution, and Social Transformation*, ed. Berch Berberoglu (Cham: Springer, 2019), 373–97; Martha McCaughey and Michael D. Ayers, *Cyberactivism: Online Activism in Theory and Practice* (New York: Routledge, 2003).

[2] Miguel Costa, Daniel Gomes, and Mário J. Silva, "The Evolution of Web Archiving," *International Journal on Digital Libraries* 18, no. 3 (2017): 191–205.

[3] Mats Berggren, "The Swedish National Digital Preservation," Riksarkivet, presentation, November 29, 2018, https://riksarkivet.se/Media/pdf-filer/dvs/RADAR_Presentation_20181129.pdf.

[4] According to some calculations, 80 percent of all web pages are altered or removed within a year, Costa et al., "Evolution of Web Archiving," 191.

[5] Lars Ilshammar, "Arkiven och den digitala paradoxen," in *Titta vad vi har! Nedslag i de enskilda arkiven*, ed. Yvonne Bergman, Barbro Eriksson, and Bo E. I. Fransson (Örebro: Folkrörelsernas arkivförbund, 2008): 247–59.

[6] See "List of Web Archiving Initiatives 2020-06-17," Wikipedia, accessed June 17, 2020, https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives.

[7] The project faced financial and judicial hindrances in 2010 and was suspended for three years; today the material is only accessible in the facilities of the Swedish Royal Library.

media platforms such as Facebook, few actors, especially among social movements, have elaborated plans for how to archive materials from their online activities (e.g., Occupy Wall Street).[8]

While working with the Swedish Protest Database in 2013, we experienced the lack of born-digital material in the Popular Movements' Archives (*Folkrörelsearkiv*) in Sweden.[9] For example, while there was comprehensive material about peace and anti–nuclear power protests from the 1960s to 1980s, there were few records of events mobilized by civil society organizations after the 1990s. As labor and social movement researchers, we realized that it was easier to do research on the 1920s than the 2010s. Even though the labor movement has great archival facilities in Sweden, investments in e-archives are costly both in terms of financial resources and in the knowledge needed to manage such systems, and thus we learned that digitalization had become a major challenge for the smaller archives in the civil society sector. Born-digital material from the Internet was not being systematically preserved. Therefore, we started the project *DigiFacket* in 2015.[10] Building solely on open-source software, we set up an automated system that collects and stores material produced by the Swedish trade union movement on the Internet (i.e., websites and social media feeds). The research team consisted of two political science researchers, one software engineer, and one research assistant. Our goal was not only to preserve the born-digital material, but also to make it easily accessible for future research projects. For this reason, early on we initiated close cooperation with the two main Swedish trade union archives—the Labour Movement's Archive and Library (LMAL) and the TAM Archives.

In this article, we describe the construction of the DigiFacket archive and discuss the lessons learned during the five years of the project. We start with a contextual description of our case selection (the Swedish trade union movement) and the existing archives of Swedish unions.

The Swedish Trade Union Movement and Its Archives

Although information on the activism of any social movement would be important to preserve, our choice to focus on the labor movement was motivated by several factors. First, the Swedish trade union movement has been the archetype of a successful trade union movement: It has exceptionally high union density and it has left a profound impact on Swedish society—particularly on the labor market and the scope of the welfare state.[11] For these reasons, the labor movement is important from historical, societal, and scholarly points of view, and preserving material from Swedish trade unions is thus of interest for many researchers in Sweden and beyond.

Second, the Swedish labor movement has a significant amount of historical material that researchers can combine with the new online material. Sweden has an established structure for archiving materials from movements and associations through the so-called "popular

[8] Anat Ben-David, "Counter-Archiving Facebook," *European Journal of Communication* 35, no. 3 (2020): 249–64; Anne Kaun, "Archiving Protest Digitally: The Temporal Regime of Immediation," *International Journal of Communication* 10 (2016): 5395–5408.

[9] Katrin Uba and Anders Westholm, "Welfare Reforms and Protest Mobilization in Sweden," working paper presented at the Swedish Sociology conference, Uppsala, April 28–29, 2016.

[10] "Digifacket," Uppsala Universitet, accessed March 9, 2020, www.statsvet.uu.se/digifacket.

[11] Jenny Jansson, "Two Branches of the Same Tree? Party-Union Links in Sweden in the 21st Century," in *Left-of-Centre Parties and Trade Unions in the Twenty-First Century*, ed. Elin H. Allern and Tim Bale (Oxford: Oxford University Press, 2017): 206–25.

movements' archives," organizations that now manage a unique and rich body of material.[12] One of these archives, the Labour Movement's Archive and Library (LMAL) in Stockholm, was established as early as 1902, shortly after the formation of the first labor organizations. This archive holds collections of all the national labor movement organizations (trade unions, labor parties, youth leagues, etc.). Today, the archive is an independent foundation overseen by the Swedish government, the Trade Union Confederation of Sweden, and the Social Democratic Party. The white-collar workers' unions founded an archive of their own in 1984: the TAM Archives, which are also located in Stockholm, and (much like LMAL) preserve all material from the national organizations of white-collar workers' unions. The TAM Archives is a meta-organization and the white-collar unions are members of the archive. The local and regional sections of all unions in Sweden are preserved by the regional popular movements' and associations' archives, which are dispersed all over the country. Thus, focusing on the labor movement is a continuation of a long tradition of preserving labor movement materials. More importantly, choosing a movement that already had its own archives has secured the long-term operation of the web archive. Although the project was initiated by scholars, our main task was to create an archiving system that could be run by the LMAL and the TAM Archives after the development phase. The archives and the trade unions themselves did not have the necessary financial resources for developing such a system in 2015 when we started the project (in fact, most of the unions had not even reflected upon the fact that their online material was not archived), but they were eager to collaborate and share their archive-specific knowledge with us.

The third reason for choosing the labor movement is technical in nature: The labor movement contains groups that are relatively easy to identify, enabling the creation of a comprehensive all-inclusive database of trade unions. Other movements, such as the environmental movement, have more blurred limits; in contrast, focusing on trade unions made it easy to identify which accounts on social media and which websites to include. Roughly speaking, the Swedish unions can be divided into four categories: the blue-collar, white-collar, upper middle class, and syndicalist unions.[13] Most trade unions are affiliated with one of the three trade union confederations: the LO (*Landsorganisationen*), the TCO (*Tjänstemännens centralorganisation*), and Saco. Only a few very small unions are not affiliated with these organizations; such unions were excluded from the project because including them would have made it difficult to find archives with the capacity or desire to continue running the web archive after the project ended. Thus, we used the LO-, TCO-, and Saco-affiliated unions as our point of departure. In total, that left us with fifty-four organizations when we started the project (three confederations and fifty-one unions). During the project, one more union became affiliated with Saco: thus, in the end, we had fifty-five organizations.

Creating the Archive

The project was never aimed to be "only" a web archive that preserves unions' born-digital material, but we, as social scientists, aimed to add a number of functions that are usually appreciated by the research community, particularly several search functions. The general technical development to date has undeniably boosted the field of digital humanities; tools for

[12] For further details, see Charlotte Hagström and Anna Ketola, *Enskilda Arkiv* (Lund: Studentlitteratur, 2018).

[13] Richard Hyman and Rebecca Gumbrell-McCormick, "Trade Unions, Politics and Parties: Is a New Configuration Possible?," *Transfer: European Review of Labour and Research* 16, no. 3 (2010): 315–31; Jenny Jansson and Katrin Uba, *Trade Unions on YouTube : Online Revitalization in Sweden* (Cham: Palgrave Pivot, 2019); Anders Kjellberg, "The Decline in Swedish Union Density since 2007," *Nordic Journal of Working Life Studies* 1, no. 1 (2011): 27.

digitization, machine learning, and big data analysis have suddenly facilitated the use of historical sources that would have taken far too long to locate before. Given this trend, we considered that researchers would want powerful tools to search material, perform content (text) analysis, and do statistical time-series analyses. These tasks require a powerful full-text search function, an indexer, and most importantly regular downloads of web pages. A regular long-term harvesting and downloading system makes it possible to create time-series data for statistical analyses and detect long-term changes, for example in the discourse of different trade unions related to employment legislation. In collaboration with the archivists from LMAL and TAM, we also agreed to set up a Wayback Machine for recreating web pages exactly as they were at specific time-points because this would benefit researchers as well as a broader set of users, such as the public or the unions themselves.

When initiating this project, a few considerations required simultaneous attention—namely, solving the ethical issues that arose and mapping the scope of the websites and social media accounts that were to be downloaded.

Ethical and Legal Considerations

Downloading and storing material from the Internet needs to be done with careful consideration. We were faced with ethical dilemmas: first, that of downloading third parties' webpages and social media feeds, and, second, the ethics of creating a database with a full-text search.

Strictly speaking, the web scraping of organizations' homepages is not illegal, and some social media companies (e.g., Twitter) provide application programming interfaces (APIs) for downloading partial feeds without the consent of the accounts' owners. However, building a *site-centric archive*—that is, the systematic and repeated scraping of a specific selection of organizations' websites—is ethically problematic, and should not be done without the consent of the creators of the websites or the owners of the accounts.[14] Furthermore, the owner of a website can detect the scraping process and block the IP address of the crawler, thus effectively disabling the harvesting process.

We solved these issues by asking for the consent of the unions to be included in the project. We created an agreement and asked the participating organizations to sign it. In the agreement, we asked for consent while simultaneously promising to install the developed software at, and hand over all material we collected during the project to, the two respective trade union archives in Sweden: the LMAL and the TAM Archives.[15] On their side, the archives promised to continue running the DigiFacket after our research project had ended. The archives were in charge of asking for consent from the unions since they already had established contacts with them. The LMAL sent out paper copies of the agreement, whereas the TAM archives emailed their unions. Almost all of the organizations signed the consent agreement; only two out of the fifty-five organizations declined our request to have their website archived. The reason for such a good turnout was our cooperation with the unions' own archives, which ensured that the unions would be in control of the material we collected. It is likely that less hierarchal organizations and social movements would not be as concerned about the systematized

[14] Julien Masanès, "Web Archiving Methods and Approaches: A Comparative Study," *Library Trends* 54, no. 1 (2005): 72–90.
[15] See LMAL at https://www.arbark.se/en/ and TAM Archives at http://www.tam-arkiv.se/.

collection of their materials as trade unions are; it is also likely that extremist groups might not agree with such a harvesting process at all.

Working with social media was much more difficult.[16] As mentioned above, Twitter allows the downloading of feeds, albeit with some limitations (for example, number of tweets, how far back in time one can scrape, and so on).[17] Facebook does not allow scraping. However, both Facebook and Twitter sell feeds. Another aspect of using APIs to download feeds was the technical side: The API changed constantly, making it a difficult task to keep up with new technical solutions. Although the project group had the technical skills to keep up with the quick pace, we knew that the archives that would be managing the system in the future did not possess such skills. Thus, in order to make the collection of social media feeds sustainable and to avoid legal issues, we had to come up with another solution. The technical, legal, and ethical difficulties in collecting social media feeds were solved by simply asking the organizations to download their accounts' histories and let us add them to DigiFacket. Both Facebook and Twitter have a download history function that the account owner can use to preserve the history of the account. The account history contains all posts made by the account (text, videos, images, and other types of media) and metadata about engagements with the post (the number of retweets, if the post was shared, etc.), but has no information created by other accounts (e.g., comments by others on the posts). Although preserving the online discussions that the unions have engaged in, including the comments made by others, has value for researchers, the benefits of our chosen solution superseded the loss of information. In order to get the unions to collect the material for us, we emailed them a request and instructions on how to download the social media histories.[18] Asking for and collecting this material was a time-consuming process for us, so we only did it once during the project. We believe that it will be easier for the archives to ask for this information, since they already ask for archival material regularly.

The second ethical dilemma relates to who has access to the information after it has been collected. Trade union membership is classified as sensitive personal data; therefore, all databases that allow full-text searches must be treated with caution. The websites are fairly unproblematic, since these are openly available and some of them can be found at the Internet Archive.[19] However, social media feeds contain more sensitive data, since they build on interaction between personal accounts. We solved part of this problem by only collecting the histories of the trade unions' social media feeds; however, this did not address all aspects of the problem. Hence, in order to ensure that DigiFacket will not be misused, the data from the websites and the data from the social media feeds are stored separately, and accessed through two different user interfaces. In order to access the social media feeds, it is necessary to have—in accordance with Swedish policy on research ethics—a research plan that has been approved by the Swedish Ethical Review Authority.[20]

Moreover, we asked the unions if they wished to impose further restrictions on the use of DigiFacket; there may be strategic reasons why the unions would not want to allow unrestricted

[16] For more on ethical dilemmas with social media research, see Hallvard Fossheim and Helene Ingierd, *Internet Research Ethics* (Oslo: Cappelen Damm akademisk, 2015).

[17] Zachary C. Steinert-Threkeld, *Twitter as Data* (Cambridge: Cambridge University Press, 2018).

[18] Our instructions can be downloaded here: "Instruktioner för nedladdning av Facebook- och Twitterhistorik," Uppsala Universitet, https://statsvet.uu.se/forskning/Fackf%C3%B6reningsr%C3%B6relsens-digitala-omvandling/publikationer/instruktioner-for-nedladdning/.

[19] Internet Archive, https://archive.org/.

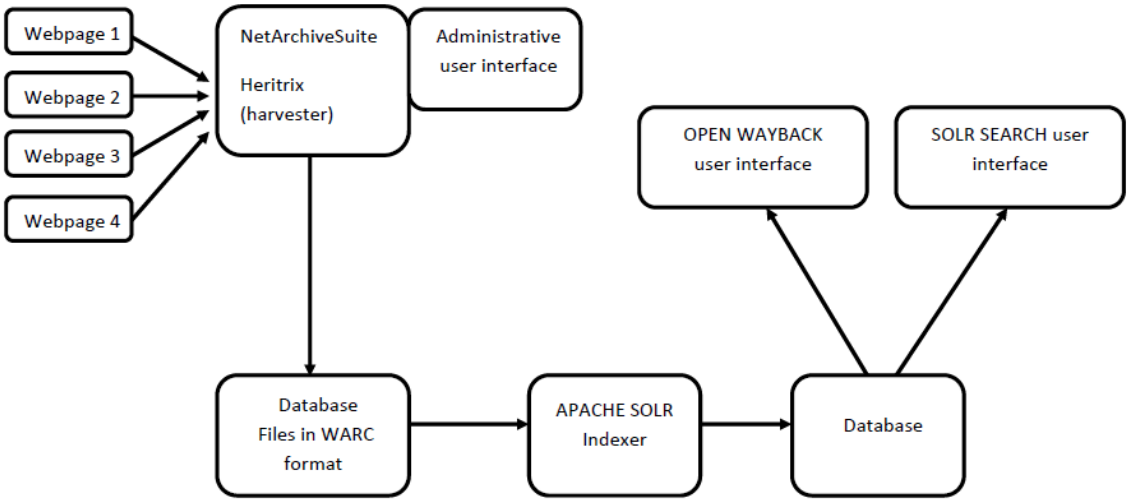
[20] See "Värnar människan i forskning," Etikprövnings myndigheten, https://etikprovningsmyndigheten.se/ for more information on that procedure.

access to their web histories (for example, radical right-wing groups trying to map the online strategies of the union movement). We decided that no further programming was needed in order to meet the demands of the unions. Thus, if further restrictions need to be applied (e.g., the user needs to apply for union permission to access DigiFacket), those will be handled by the LMAL and TAM Archives once we have transferred the system to them.

Structure of DigiFacket

DigiFacket is made up of several different components. The downloading cycle roughly consists of the following steps: harvesting, storing, indexing, and displaying the material for the user. These steps are displayed in figure 1, and will be further explained below.[21]

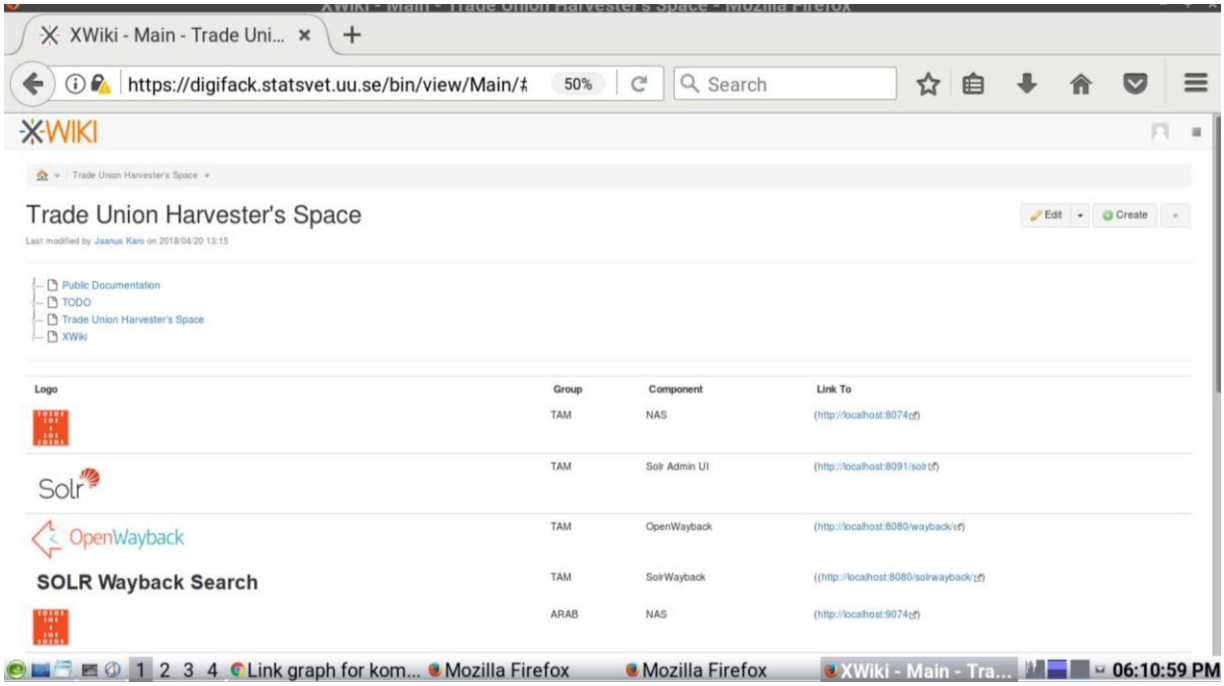
Figure 1. The components of the DigiFacket system



Since several people were working on the project and several software components were operating simultaneously, we built a closed wiki page through which all interfaces were accessed (figure 2). The wiki also enabled us to preserve information about all of the steps of the working process and to exchange information within the project group.

[21] For full information on the technical aspects, see our technical report at the project's webpage.

Figure 2. Screenshot of the wiki page for accessing all of DigiFacket's components



Step 1: Harvesting

Choosing the labor movement as our focus made it easy for us to proceed with the very first task in the process of building the archive: mapping which Universal Resource Locators (URLs) to download. Since we had complete lists of the unions' names, the official websites of the unions were easy to identify. We also searched for other websites run by the unions, such as temporary campaign pages, and for websites run by side organizations, such as specific youth organizations or local sections. In total, we ended up with approximately one hundred web pages.

After identifying which URLs to download, the next task was to decide whether to write suitable software from scratch or use available software packages. Upon browsing the available freeware for web scraping, we quickly realized that there were several good options. The web-archiving community has developed a number of software programs that function well.[22] Thus, we tested several software programs before settling on NetarchiveSuite (NAS) and the data crawler Heritrix. The software we tested included curl, wget, HTTrack, and Norconex HTTP Collector.[23] All had advantages and disadvantages, but the Heritrix is today the most popular crawler for web archives.[24]

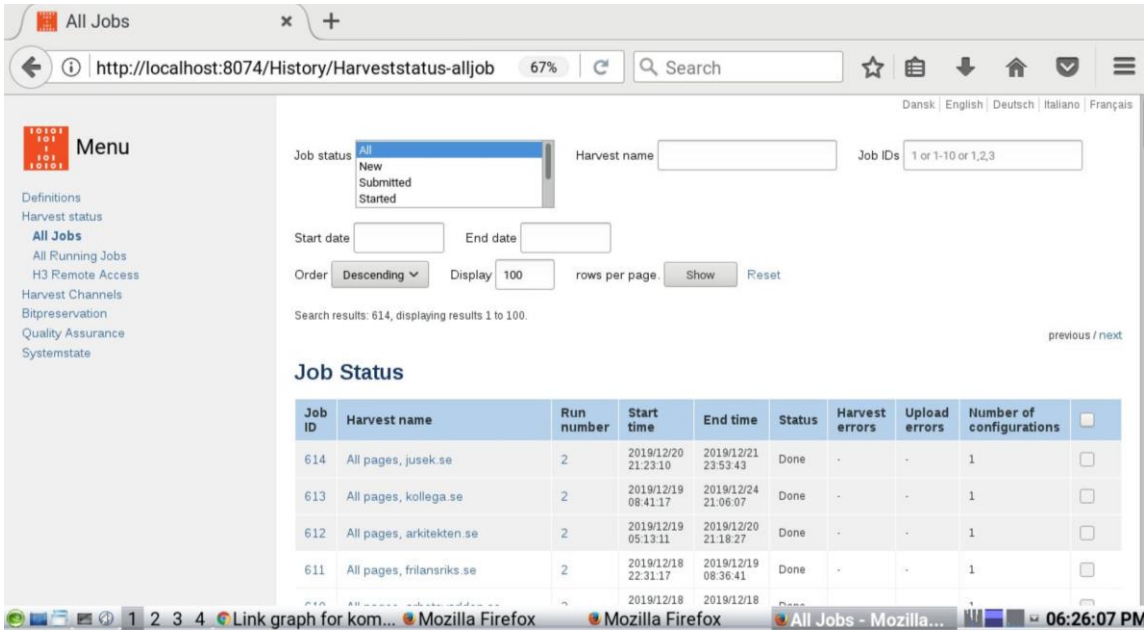
Our first complete harvesting cycle, in which we tested downloading the websites of all of the organizations, was done with HTTrack. As it is a commonly used web-archiving system,

[22] Muzammil Khan and Arif Ur Rahman, "A Systematic Approach towards Web Preservation," *Information Technology and Libraries* 38, no. 1 (2019): 71–90.
[23] "curl," accessed November 22, 2019, https://curl.haxx.se/; "WGet," GNU Operation System, accessed November 22, 2019, https://www.gnu.org/software/wget/; "HTTrack Website Copier," accessed November 22, 2019, https://www.httrack.com/; "Norconex HTTP Collector," accessed November 22, 2019, https://www.norconex.com/collectors/collector-http/.
[24] Emily Maemura et al., "If These Crawls Could Talk: Studying and Documenting Web Archives Provenance," *Journal of the Association for Information Science and Technology* 69, no. 10 (2018): 1223–33.

HTTrack worked well for scraping websites; however, we encountered other problems.[25] Since DigiFacket was never meant to merely preserve websites, but was also intended to make the data easy for researchers to access and to provide numerous options for searching the information extracted from the web pages, we needed to connect the scraped web pages with an indexer; however, we had problems finding a powerful indexer that worked well with HTTrack. As a result, NAS turned out to be the preferable choice for us. Developed by the Danish National Library, NAS has become a leading software for scraping websites.[26] This software suite was developed for archiving web pages with index and search functions, so it met the functional requirements of our project.

Through the NAS admin user interface (UI) component, curators can easily add and remove URLs (see figure 3), and can configure the scraping (for example, time intervals, number of hops [subpages], maximum number of gigabytes [GB] per URL, etc.). Based on these configurations, this software component collects web data and wraps it up into a database. The NAS UI enables the curator to monitor the harvesting process, which must be done regularly. If something is not working with the harvesting, the curator is notified through the NAS UI.

Figure 3. Screenshot of the harvesting process via NetarchiveSuite



Since we knew that the material we collected while building DigiFacket would be distributed to two different archives, we set up two parallel systems from the start: one for collecting websites from the LO-affiliated unions and the syndicalist movement (to be handed over to the LMAL) and one for collecting websites from the TCO- and Saco-affiliated unions (to be delivered to the TAM Archives). NAS allows the user to choose the number of crawlers, so it is possible to run parallel sessions. We chose to use four parallel sessions: two for each archive.

[25] See "List of Web Archiving Initiatives," accessed March 9, 2020, https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives.
[26] Ditte Laursen and Per Møldrup-Dalum, "Netarkivet 10 år," *Revy* 39, no. 2 (2016): 6–9.

Originally, we set out to download the first page of each website once a week and the entire website once a month. After a number of downloading cycles, we modified this configuration. This was due to several factors.

First, one cycle must be finished before the next can begin; thus, the speed of the harvesting process had to be taken into consideration. The speed depends not only on the Internet connection of the crawler, but also on how it affects the server of the website that it downloads. High-speed downloading can make other visitors perceive the web page as slow, and we did not want to interfere with the traffic on the unions' websites. Instead, we tried to be as "polite" as possible while harvesting. We set the crawler used by NAS (i.e., Heritrix3) to a speed of 50 kilobytes per second (KB/s), which was equivalent to the usage of a normal visitor on the website. However, at this speed, the cycle took quite a long time. It is important to note that websites contain much more information than just the text; over time, websites have become increasingly complex and now consist of layers of different functions, images, videos, moving elements, and so forth. All these functions generate a large amount of metadata (information about the web page's construction), which slows down the harvesting process. The websites in our sample range from 100 megabytes (MB) to 5 GB in size; consequently, the time needed to harvest a whole website varied from a few hours to several days.

Second, we estimated how much information would be lost if we downloaded the entire website less frequently than once a month. The critical part of any web-archiving system is to capture information that is posted on a web page for a short amount of time—information that would quickly be "lost." We assumed that most of that type of information would be posted on the first page of the unions' websites, but in order to make an accurate configuration of the downloading frequency, we monitored how much changed on the unions' subpages over several weeks. Our conclusion was that although most of the subpages remained unchanged, the first page changed often—for the big unions, almost on a daily basis. Moreover, the changes we could detect on the subpages were mostly added information (e.g., information that was initially posted on the first page was later moved to a subpage, instead of being deleted). The information posted on subpages was seldom removed. Thus, we concluded that the subpages could be downloaded with a larger time gap without too much information being lost. However, it was very important to download the first page frequently. For this reason, we set one crawler per archive to download the whole of each website once every second month and one crawler to harvest the first (main) pages once a week.

Step 2: Storing

To facilitate the sustainable, long-term storage of the material, the files needed to be compressed. Every website in our sample contained thousands of files on average, so frequent harvesting rapidly created a large number of files. One important advantage of choosing the NAS software package was that it supported the Web ARChive (WARC) file format. For each harvesting cycle, NAS gathers all harvested files into a single WARC file, which simplifies data transferring, indexing, storing, and other processes. The WARC format specifies a method for combining multiple digital resources into an aggregate archive file together with related information. As an ISO standard (ISO 28500:2017), WARC is recognized by most national library systems as the standard to follow for web archiving and is used by many of the software programs that manipulate harvested data. It is likely to be a format that will be compatible with other software for the foreseeable future.

During the project, we collected material comprising 4 terabytes (TB) of data, which included the metadata containing information about the harvesting process. Hence, if necessary, it would be possible to use the space more efficiently and remove this metadata without any implications for the UIs, Wayback Machine, or Apache Solr search.

Steps 3 and 4: Indexing and Displaying through User Interfaces

Since we considered it to be vitally important to enable full-text searches of all the data, we needed software for indexing. Full-text searching is often the preferable option for researchers. Nevertheless, because the DigiFacket material is large, quickly growing, and contains duplicates (i.e., the same files have been repeatedly downloaded), one search may result in a very large number of hits. For example, searching for "work environment" generates thousands of hits, many of which are duplicates. Of course, this makes it difficult to efficiently search the material. Indexing is one way of solving this problem. Adding specific delimitations to the searches, such as time frame and domain name, helps to narrow down the number of results.

Although NAS has an indexer of its own, it appeared to be too rigid to allow us to adjust the system according to our needs, especially in terms of the possibility of adding our own search categories to the UI. Instead, we tried out and decided to go with a combination of SolrWayback and Apache Solr; the latter is a search platform written in Java and developed by the Apache Lucene project.[27] Solr permits the indexing of large amounts of material, and a converter from WARC files is available.[28] Our first harvesting cycle was downloaded using HTTrack and was thus not stored in WARC. Unfortunately, we discovered that this material could not be automatically indexed by Solr. Thus, it was not possible to do a full-text search of the material starting from cycle one, nor was it possible to open the data in the Wayback Machine. The files from HTTrack could only be browsed by a local web browser. Thus, although the retrieved data was not lost, there are limits to how this particular data can be used. This error of ours highlights the importance of careful planning. The aim of the archive and the possible needs of future users must be clear before deciding on technical solutions.

DigiFacket has two UIs: one to browse previously archived web materials—called OpenWayback—and one for letting users make full-text search queries within the same data. For the latter, a combination of Apache Solr and SolrWayback is used.

OpenWayback is an open-source Java application. Its predecessor was first released by the Internet Archive in September 2005, based on the Internet Archive Wayback Machine, to enable public distribution of the application.[29] It was called the Open-Source Wayback Machine (OSWM). The Internet Archive handed over the lead repository of the OSWM to the International Internet Preservation Consortium (IIPC), which launched the OpenWayback project in October 2013. The Internet Archive continues to develop Wayback. The plan is to include and merge changes on the Internet Archive fork so that the two do not diverge significantly. Thus, this software solution will be updated and developed, which means that the DigiFacket system also has the potential to follow the rapid evolution of the Internet.

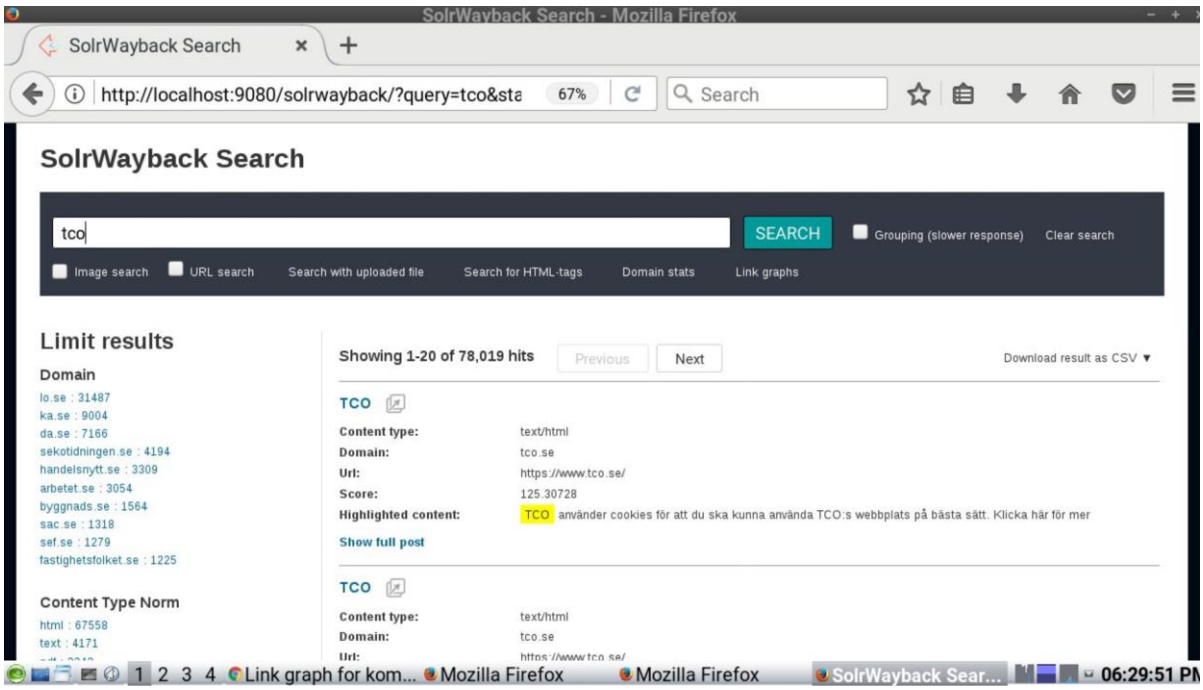
[27] David Smiley et al., *Apache Solr Enterprise Search Server* (Birmingham: Packt Publishing Ltd., 2015).
[28] "Webarchive-discovery," github.com, accessed March 9, 2020, https://github.com/ukwa/webarchive-discovery/tree/master/warc-indexer.
[29] Internet Archive, accessed March 9, 2020, https://archive.org/web/web.php.

SolrWayback is not only a web application for browsing previously harvested WARC files (similar to OpenWayback), but also a full-text search interface built on top of the Apache Solr server, where WARC files have been indexed using the British Library WARC Indexer.[30] In addition to its powerful Lucene-like query syntax, SolrWayback can be used to search images, export search results to a WARC file, and more. Before settling on SolrWayback, we tested another UI: Shine.[31] Shine is a prototype web archive exploration UI developed by the British Library; it is also based on the Apache Solr back end and the WARC Indexer.[32] However, we soon discovered that Shine was lacking several types of functionality that we needed in DigiFacket, which had already been implemented in the SolrWayback code base.

Making use of descriptive, structural, and administrative metadata information, the SolrWayback search has a number of predefined categories that help to sort the hits: type of content (file type), domain, year of download, and so forth.[33] Combining these with the full-text search refines the search function. In addition to these categories, it is possible to add more categories through the construction of a thesaurus. From the start, we had great ambitions to make an index through an elaborated thesaurus. However, since the search syntax of Apache Solr is efficient, and the techniques and software for the content analysis of large-text materials have developed quickly, the need for such an index seemed less and less relevant. Moreover, it turned out to be very difficult to create an efficient thesaurus, since the content of the unions' websites is similar for all of the organizations in our sample: For example, all of the unions write about "work environment issues" on their websites, and the websites are repeatedly downloaded. It turned out to be very difficult to create a thesaurus that would add value to the full-text search. We do not rule out the possibility of a thesaurus being constructed in the future; technically, a thesaurus could be added to the Solr search. However, we chose to leave this task to the professional archivists and/or librarians at the LMAL and the TAM Archives.

[30] "SolrWayback," github.com, accessed March 9, 2020, https://github.com/netarchivesuite/solrwayback.
[31] "Shine," github.com, accessed March 9, 2020, https://github.com/ukwa/shine.
[32] "Webarchive-discovery," accessed March 9, 2020, https://github.com/ukwa/webarchive-discovery.
[33] Khan and Rahman, "A Systematic Approach."

Figure 4. Screenshot of the Solr search system adapted for DigiFacket



As mentioned above, websites contain much more information than just the text. In fact, text is the easy part to search for: indexing images and videos is harder. In our system, we only use the metadata to search for non-text objects. Image analysis would require scripts to automatically identify objects in images, which we have not been able to add to the system as yet (although this software could be added in the future). Analyzing images and videos manually is not feasible. The difficulty of making images and videos searchable in web archives highlights the importance of creating good metadata when creating a web page.

Lessons Learned

Building the DigiFacket archive has been a learning experience for the whole project team. Seeing the outcome of the project, we believe that the archive will be of use for a diverse set of researchers from different disciplines: the archive can be used to study trade union–related research questions such as collective bargaining, international cooperation, how the recent Covid-19 virus outbreak was handled by the unions, what happens with the relations between different unions during the mobilization of strikes and protests, and so on. The archive can also be used to study the technical development of websites such as design, Java programming, the use of different file formats, and more.

In our experience with the DigiFacket archive we learned some specific lessons:

(1) Although digitization has challenged the archival sector, the general development of information technology has offered solutions in terms of new possibilities for storing and indexing large amounts of material. The potential for archiving and providing access to archival materials is endless with the right software.[34] During our work with DigiFacket, we understood that technology changes very quickly and this rapid development creates a number of dilemmas

[34] Colin Post, "Building a Living, Breathing Archive: A Review of Appraisal Theories and Approaches for Web Archives," *Preservation, Digital Technology & Culture* 46, no. 2 (2017): 69–77.

for harvesting and storing born-digital material. For instance, web-page structures become more complicated over time and one has to adjust harvesting and storing processes frequently. There is also a need to understand the process' log files, which help to detect and fix the problems before any data is lost. Therefore it seems that close cooperation between archives and software engineers is needed. Ultimately, web archiving would benefit from archivists with programming skills and ideally, programming would be part of archival education. In our case, the archivists received a short training for handling smaller problems such as changing the web addresses to be harvested, while the more complicated tasks must still be completed by the software developer.

(2) Indexing this large amount of material proved to be very difficult. Unsurprisingly, we ended up relying on metadata to create search categories. The metadata in our sample, which here mainly means different tags with keywords given to audio files, videos, and photos, was of very diverse quality, perhaps because the creators of the information (trade unions) gave no thought to metadata. In our specific case, the unions' archives could actually provide the unions with instructions or recommendations on how to create useful metadata (e.g., recommendations on how to best construct alternative texts ["alt tags"], how to name images and pdf files, etc.)—just like they have given unions instructions for the sorting of regular paper-based materials. Archiving paper documents is well embedded in the labor movement culture in Sweden, however the unions' web pages and social media feeds are often handled by professional public relations and communication officers that may not have a background in the movement. Thus, for preservation of the born-digital materials from the Internet, instructions from the archives may be even more important than for traditional materials.

(3) Web archiving for smaller institutions and organizations has much greater flexibility than relying solely on large web-archiving projects, such as the Internet Archive. For the Swedish trade unions that already have their own archives, it makes sense for the born-digital material they produce to end up in their own archives, where it complements traditional organizational materials. Moreover, organizations have control over how the material in their own archives is stored and used. In contrast, relying on private enterprises such as Facebook to preserve and make different accounts' histories available is not a secure method: private enterprises can easily change the user terms and delimit such possibilities. Therefore, we encourage all archives preserving material from social movements and organizations with social media accounts to ask their organizations to download their social media feeds and share them with the archive. This is a completely legal way to preserve an increasingly important part of organizations' and movements' communication.

(4) Although archivists are certainly well trained to identify the digital materials to preserve, it is useful for scholars using the archive and archivists to cooperate.[35] Cooperation could help to define the scope of the material to preserve, the types of platforms to focus on, and the frequency and depth of website crawling. We started to use the collected material, particularly the YouTube videos posted by trade unions, at an early stage of the project in order to keep the user's perspective when developing the archive.[36] Being interested in the message of the videos rather than their visual characters, the text—the metadata of titles and descriptions provided by the unions—that was withdrawn from the video material became the most important source of information. Other scholars might be more interested in the music accompanying the videos or the specific visual genre of the videos. The nature of the research always dictates which kind

[35] Post, "Building a Living, Breathing Archive."
[36] See more in Jansson and Uba, *Trade Unions on YouTube*.

of material is needed and ideally one would download everything as frequently as possible. However, in the context of limited resources there is a need for setting priorities and an ongoing dialogue between scholars and archivists would be particularly useful in such situations.

Bibliography

Ben-David, Anat. "Counter-Archiving Facebook." *European Journal of Communication* 35, no. 3 (2020): 249–64, DOI: 0267323120922069.

Carty, Victoria, and Francisco G. Reynoso Barron. "Social Movements and New Technology: The Dynamics of Cyber Activism in the Digital Age." In *The Palgrave Handbook of Social Movements, Revolution, and Social Transformation*, edited by Berch Berberoglu, 373–97. Cham: Springer, 2019.

Costa, Miguel, Daniel Gomes, and Mário J. Silva. "The Evolution of Web Archiving." *International Journal on Digital Libraries* 18, no. 3 (September 2017): 191–205.

Fossheim, H., and H. Ingierd. *Internet Research Ethics*. Olso: Cappelen Damm akademisk, 2015.

Hagström, Charlotte, and Anna Ketola. *Enskilda arkiv*. Lund: Studentlitteratur, 2018.

Hyman, Richard, and Rebecca Gumbrell-McCormick. "Trade Unions, Politics and Parties: Is a New Configuration Possible?" *Transfer: European Review of Labour and Research* 16, no. 3 (2010): 315–31.

Ilshammar, Lars. "Arkiven och den digitala paradoxen." In *Titta vad vi har! Nedslag i de enskilda arkiven*, edited by Yvonne Bergman, Barbro Eriksson, and Bo E. I. Fransson, 247–59. Örebro: Folkrörelsernas arkivförbund, 2008.

Jansson, Jenny. "Two Branches of the Same Tree? Party-Union Links in Sweden in the 21st Century." In *Left-of-Centre Parties and Trade Unions in the Twenty-First Century*, edited by Elin H Allern and Tim Bale, 206–25. Oxford: Oxford University Press, 2017.

Jansson, Jenny, and Katrin Uba. *Trade Unions on YouTube : Online Revitalization in Sweden*. Cham: Palgrave Pivot, 2019.

Kaun, Anne. "Archiving Protest Digitally: The Temporal Regime of Immediation." *International Journal of Communication* 10 (2016): 5395–5408.

Khan, Muzammil, and Arif Ur Rahman. "A Systematic Approach towards Web Preservation." *Information Technology and Libraries* 38, no. 1 (2019): 71–90.

Kjellberg, Anders. "The Decline in Swedish Union Density since 2007." *Nordic Journal of Working Life Studies* 1, no. 1 (2011): 27.

Laursen, Ditte, and Per Møldrup-Dalum. "Netarkivet 10 år." *Revy* 39, no. 2 (2016): 6–9.

Maemura, Emily, Nicholas Worby, Ian Milligan, and Christoph Becker. "If These Crawls Could Talk: Studying and Documenting Web Archives Provenance." *Journal of the Association for Information Science and Technology* 69, no. 10 (2018): 1223–33.

Masanès, Julien. "Web Archiving Methods and Approaches: A Comparative Study." *Library Trends* 54, no. 1 (2005): 72–90.

McCaughey, Martha, and Michael D. Ayers. *Cyberactivism: Online Activism in Theory and Practice*. New York: Routledge, 2003.

Post, Colin. "Building a Living, Breathing Archive: A Review of Appraisal Theories and Approaches for Web Archives." *Preservation, Digital Technology & Culture* 46, no. 2 (2017): 69–77.

Smiley, David, Eric Pugh, Kranti Parisa, and Matt Mitchell. *Apache Solr Enterprise Search Server*. Birmingham: Packt Publishing Ltd., 2015.

Steinert-Threkeld, Zachary C. *Twitter as Data*. Cambridge: Cambridge University Press, 2018.