Internet Histories Digital Technology, Culture and Society Internet histories and computational methods: a ”round-doc” discussion

Niels Brügger, Ian Milligan, Anat Ben-David, Sophie Gebeil, Federico Nanni, Richard Rogers, William J Turkel, Matthew S Weber, Peter Webster

Internet Histories
Digital Technology, Culture and Society

ISSN: 2470-1475 (Print) 2470-1483 (Online) Journal homepage: https://www.tandfonline.com/loi/rint20

Internet histories and computational methods: a "round-doc" discussion

Niels Brügger, Ian Milligan, Anat Ben-David, Sophie Gebeil, Federico Nanni, Richard Rogers, William J. Turkel, Matthew S. Weber & Peter Webster

To cite this article: Niels Brügger, Ian Milligan, Anat Ben-David, Sophie Gebeil, Federico Nanni, Richard Rogers, William J. Turkel, Matthew S. Weber & Peter Webster (2019): Internet histories and computational methods: a "round-doc" discussion, Internet Histories, DOI: 10.1080/24701475.2019.1639352

To link to this article: https://doi.org/10.1080/24701475.2019.1639352

Published online: 29 Aug 2019.

INTERNET HISTORIES
https://doi.org/10.1080/24701475.2019.1639352

Internet histories and computational methods: a "round-doc" discussion

Niels Brügger[a], Ian Milligan[b], Anat Ben-David[c], Sophie Gebeil[d], Federico Nanni[e], Richard Rogers[f], William J. Turkel[g], Matthew S. Weber[h] and Peter Webster[i]

[a]School of Communication and Culture, Aarhus University, Aarhus, Denmark; [b]Department of History, University of Waterloo, Waterloo, Canada; [c]The Department of Sociology, Political Science and Communication, the Open University of Israel, Ra'anana, Israel; [d]the Telemme Laboratory, Aix-Marseille University, Marseille, France; [e]Data and Web Science Group, University of Mannheim, Mannheim, Germany; [f]Media Studies, University of Amsterdam, Amsterdam, the Netherlands; [g]Department of History, The University of Western Ontario, London, Ontario, Canada; [h]Hubbard School of Journalism and Mass Communication, University of Minnesota, Minneapolis, MN, United States; [i]Webster Research and Consulting Ltd, Chichester, UK

ABSTRACT

This conversation brings together leading experts within the field of computational methods. Participants were invited to discuss "Internet histories and computational methods", and the debate focused on issues such as why scholars of internet histories should consider using computational methods, what scholars should be looking out for when they use these methods, how the process of collecting influences computational research, what impedes the use of computational methods, to what an extent internet historians should learn to code (or conversely, if developers should learn about historical methods), what are the most defining moments in the history of computational methods, and, finally, the future of using computational methods for historical studies of the internet.

ARTICLE HISTORY
Accepted 1 July 2019

KEYWORDS
computational methods; digital humanities; digital methods; web; internet; history

As editors of the "Internet histories and computational methods" special issue of *Internet Histories*, we wanted to bring together leading experts to participate in a roundtable discussion. Yet the difficulty of physically bringing together people was considerable, and we decided that it would be timely and appropriate to use the affordances of the internet: the ability to collaborate remotely! Experts were thus invited to participate in what we called a "Round-Doc" discussion, i.e. a roundtable conversation taking "place" not in a physical room, but rather in a shared Google document.

Although perhaps less ambitious, our intent was directly linked to one of the first collaborative writing experiments which took place on a computer network, namely the "Epreuves d'écriture", "Writing Samples", that were part of the exhibition *Les Immatériaux*, organised by French philosophers Jean-François Lyotard and Thierry Chaput in 1985 at Centre Georges Pompidou in Paris. For the exhibition, the organisers invited around 30 authors, writers, scientists, artists, philosophers and linguists to comment on a list of 50

words related to the topic of the exhibition, the immaterial(s). As this was in the very early days of computer networks, the organisers had to supply sponsored micro-computers and networks (cf. Lyotard & Chaput, 1985).

Much has happened with the internet since 1985, and we are now able to make a similar, yet much smaller, collective writing experience in a much easier way. The topic of our discussion here is "Internet histories and computational methods", and the process took place in the following way. First, we posed seven questions and seven scholars with a strong track record within the field were invited to respond. Then the document was opened for comments from two scholars each for one week, and after one week it was closed for these participants and opened up for two other contributors. Once all contributors had participated the document was open to all for two weeks. During this last stage, a couple of new questions were added. Finally, the document was edited.

The initial questions were as follows:

- Why should scholars consider using computational methods when they study the many forms of the internet, including the web and social media? Can you illustrate this by one or more examples?
- What should scholars be looking out for when they use these methods? What are the possible pitfalls and challenges?
- To use computational methods, the object of study needs to be in digital form. Do you have any thoughts about to what extent the process of collecting influences computational research? Are the right sources collected? In the right format? By the right institutions?
- Is there anything that in your mind impedes the use of computational methods in studies of the history of the internet? Are source collections not "researcher-friendly?" Is there a lack of adequate methods and tools? Are there other obstacles?
- How do you see the relation between subject-matter experts like historians and new media scholars and developers (from systems librarians to programmers)? Should internet historians learn to code, or conversely, is the onus on developers to learn about historical methods?
- The use of computational methods in historical study has a history of its own. What are the most defining moments in the history of computational methods?
- How do you see the future of using computational methods for historical studies of the internet? What are the biggest challenges? The biggest opportunities or most exciting projects today? Which type of methods and tools would you like to see developed?

The following scholars participated in the conversation:

- Anat Ben-David, Ph.D., senior lecturer in the Department of Sociology, Political Science and Communication at the Open University of Israel.
- Sophie Gebeil, Ph.D., senior lecturer in contemporary history at the Telemme Laboratory, Aix-Marseille University.
- Federico Nanni, Ph.D., Postdoctoral Researcher, Data and Web Science Group, University of Mannheim.

- Richard Rogers, Ph.D., Professor, Chair in New Media & Digital Culture, University of Amsterdam.
- William J. Turkel, Ph.D., Professor of History at The University of Western Ontario.
- Matthew Weber, Ph.D., Associate Professor, Hubbard School of Journalism and Mass Communication, University of Minnesota.
- Peter Webster, Ph.D., independent scholar and consultant, Webster Research and Consulting Ltd, UK.

We hope you enjoy this "round-doc" as much as we had convening it!

*

Niels Brügger and Ian Milligan: Why should scholars consider using computational methods when they study the many forms of the internet, including the web and social media? Can you illustrate this by one or more examples?

Sophie Gebeil: As a historian, I focus on how internet users deal with the past, memory, or heritage in general in the 2000s. As such, the traces of the web are considered as born-digital material, stabilised thanks to the process of archiving the web (reborn digital material). Even if in my research, I have favoured a micro-historical and qualitative approach, I think that as a historian who exploits online sources, we cannot ignore the question of computational methods. It intervenes at different stages of the research. When the corpus is compiled, they make it possible to identify dynamics and trends through data analysis. When analysing the corpus once it has been circumscribed: even working from the different archived versions of a single website such as histoire-immigration.fr, the amount of data available exceeds the researcher's capacity. However, the digital nature of Web traces makes it possible to begin to explore certain issues such as the evolution of the lexicon mobilised on the site (plain text) or the outgoing URLs in the different versions. For instance, in case of the history-immigration.fr host, we can see the data available for the period 2003–2019 on the Internet Archive.[1] This offers potential for analysis that is complementary to traditional historiographical methods.

Anat Ben-David: The internet is a computational medium, so it is rather tempting to apply computational methods in internet research, for there seems to be a convenient "structural fit" between medium and method. In particular, computational methods are helpful when researchers are interested in processing and analysing large quantities of data that cannot be processed manually or qualitatively. For example, it would be useful to use computational methods if one is interested in studying the evolution of hyperlink networks over time, or in analysing millions of tweets around a certain event; in finding patterns in internet traffic data, or in comparing user profiles across social media platforms. However the temptation to use computational methods in internet research just because of the "structural fit" might be misleading, as these methods may be helpful in answering certain research questions, but not others. For example, in Web history, computational methods are very useful in extending the scope of analysis that is often limited by interfaces to web archives, allowing to perform "distant reading" of entire national webs. However these meta-level analyses are useful in describing the structure of networks, or in detecting patterns in content, but they might fail to answer historical questions about user experiences and motivations, or in describing historical narratives in detail.

Federico Nanni: I think your point about a "structural fit" is a central point, using (especially shiny advanced) computational methods is often tempting, but our research should be guided by a specific question, not by a fascinating methodological application.

William J. Turkel: To take the opposite perspective on what guides our research, I think that many interesting questions can emerge when you start with a particular computational method or approach and think of ways to apply it to historical sources. For example, someone who is broadly interested in geospatial phenomena may decide to learn GIS. As their knowledge deepens they begin to find uses for techniques that they did not originally know about. They become able to read literature that was previously inaccessible, thus becoming acquainted with new research questions. This is one established career path in science, engineering or applied math: become an expert in digital signal processing, or nonlinear optimisation, or differential equations, then work on problems that suit your tools. At the most general level, I would argue that learning how to program is a similar strategy. In the case of working with born-digital sources, you can be confident your programming skills will not go to waste, even if you do not start with a specific set of questions that you know you will be able to answer.

Federico Nanni: I agree with what is said above by Sophie and Anat, but at the same time, I think it's necessary to distinguish between the use of computational methods in two different broad areas, the first related to the retrieval of information and the second to the quantification of information. While they are clearly interconnected (an information retrieval tool, for instance, based upon query-likelihood, relies on quantitative assumptions) the researcher's final goals are different. Dealing with large-scale web collections, information retrieval tools are almost always needed, at least as an access point, even when we plan to conduct a specifically focused qualitative analysis. It is therefore important that researchers understand what type of results such tools provide (and especially not provide) when we enter a query, starting from the basic fact that when employing many of these tools (e.g. Twitter search), results are tailored around our interests, location, previous interactions, etc.

Conversely, when researchers adopt computational methods for capturing and modeling a quantitative property of the data under study, they are entering into a very different framework of analysis, strongly aligned with social science approaches.[2] If we consider for instance conducting a diachronic sentiment analysis of a political campaign from social media data, researchers should work under the assumption that a specific tweet collection is a representative sample of the population under study, that a particular phenomenon (stating a positive/negative opinion towards a candidate) can be objectively measured and precisely captured with a computational approach (e.g. a dictionary-based sentiment classifier), that this can be used as a proxy for a real-world behaviour (e.g. being pro/against a candidate), and so on.

Richard Rogers: Instead of thinking of methods as toolboxes to be trucked to the next field of inquiry, no matter its distance, one could argue that media historically have had their classic methods for their study such as how studies of television prefer reception and film spectatorship. In European media studies, screen studies may be imported fruitfully into internet studies (such as encoding and decoding), but the

internet could be said to be (among others) a medium of algorithmic recommendation. It would follow from a "methods of the medium" outlook to consider both critically unboxing these systems as well as applying them for the study of the workings of the internet. Interrogations into polarisation, the filter bubble, so-called fake news and other recent preoccupations in internet studies would benefit from capturing the hierarchies of people, stories and sources created to order media objects for the user. Studying how principles of homophily often override heterophily when understanding user preference presupposes an inquiry into how machines are trained to learn. In both these instances, one employs computational techniques but is not restricting method to the computational; rather the computational is part of the critical repertoire.

Matthew Weber: My colleagues have raised a number of important points in thinking about why scholars should use computational methods when studying questions pertaining to the internet, the web and social media. Computational methods are central in my research as a scholar focusing on digital technology and studying processes of technology evolution and change. I agree with much of what has been said; different domains of research have different traditions with regard to methods, information retrieval and quantification both often call for computational methods in this space, and often there is a structural fit.

That said, when I speak to graduate students and others who are just working their way into this space, I first ask about their research questions and the theory guiding their work. I rarely think the selection of the method should drive the research; William raises some important points about the power of computational methods to guide exploration, but from a social science perspective the structural fit of the method will generally trump exploration. To that end, I generally believe research questions should be derived from prior research, from personal motivations, and from an exploration of emerging patterns and new trends in society.

In the context of internet studies and historical studies, it is increasingly the case that questions focused on the web require the researcher to sift through data on a scale not seen in prior generations. Thus, this research questions in this domain often push scholars to work with computational methods. Personally, I came into this domain of research wanting to study how news media organisations were adapting to web technology in the early 2000s. I had read enough small-scale case studies to know that I wanted to understand broader patterns; my research questions, grounded in organisational studies, led me to computational methods because of the scope and scale of the questions that I was asking.

Peter Webster: One of the more sterile tropes of much of the discussion about the digital humanities is an opposing of "traditional" and "digital" methods, as if it were necessary that one of the two should be all-sufficient. This has much to do with scholarly politics and speculation about where the next piece of research funding will most likely be directed, but it is clear to me that digital methods in general add possibility – to answer the questions that could not feasibly be approached before – and that there is no reason to suppose that traditional historical method is thereby somehow under threat of extinction. Nothing in the fact of distant reading prevents me from

also reading closely. As such, if a scholar sees a research question that can only be answered by the adoption of a computational approach, then that is the approach to take. That should not stop the deployment of traditional methods alongside it, and (as has often been the case in my own work) what tends to occur is a tracking back and forward between methods as the inquiry possesses. My computational processing of link graphs has often begun with a small number of known domains, and leads to the close reading of others, and then in turn to more interrogation of the graph. It is this integration at which we should be aiming.

Picking up the question of which comes first, method or question, it is surely the case that both are true. I could hardly even fire up a particular application without having some initial question to ask. Over time, my research question will evolve as the work progresses. And (as William suggests) I will most likely emerge at the end with new questions to pursue, some of which have only occurred to me as a result of coming to know more about the method. The two cannot meaningfully be untangled.

Niels Brügger and Ian Milligan: What should scholars be looking out for when they use these methods? What are the possible pitfalls and challenges?

Sophie Gebeil: There are many traps, individually and collectively. This depends a lot on the researcher's discipline and their own computer skills. In the case of historians, their computer skills are mostly underdeveloped in initial training. In this case, I identify two pitfalls related to their relationship to technology. First, historians' fascination with the belief that the computational tool is inherently objective, thus reconnecting with the myths of quantitative history and abandoning disciplinary hermeneutics by reducing the understanding of human societies to data alone. Second, the lack of interest that would amount to entrusting the application of the method to a data scientist without taking into account calculation biases in the analysis. On the other hand, by participating and cooperating in the choice and development of the analytical method, there are many possibilities. What interests me most is the possibility of monitoring the evolution of online content over time, which implies designing tools dedicated to web archives.

Anat Ben-David: In my view, a useful guiding question could be whether or not there is added value in introducing computation to the analysis. For example, when conducting content analysis of large volumes of text, researchers often use off-the-shelf methods, tools and scripts such as sentiment analysis, or topic modelling, that provide statistically based summarisation and classification of text. As these off-the-shelf methods gain popularity, they tend to become black-boxed (in that the user doesn't understand the theoretical justifications, histories and limits of the underlying method) and lose their critical edge. However, when computational tools and methods are critically devised and specifically tailored to answer specific research questions, they open up a variety of exciting new ways of thinking about research questions, and of answering them creatively, reflexively and critically.

An anecdote from one of my previous research projects illustrates both the potential and limits of the computational approach to internet research. My colleagues and

I were interested in characterising the typical colors of national webs, and to measure the "distance" between the average pallet of web pages from the color of the national flag (Ben-David, Amram, & Bekkerman, 2018). We used an off-the-shelf technique called K-means clustering to create average monthly color histograms of the entire national web of the former Yugoslavia. We found that the similarity of the domain's colors to the colors of the Yugoslav national flag decreases over time. However, qualitative analysis of a sample of the analysed images revealed that many of them displayed a tiny flag at the corner of the image. While the computational method allowed us to summarise the colors of about 40,000 Web pages, it was blind to the symbolic presence of these miniature flags.

William J. Turkel: One of the most valuable outcomes of working with specifically tailored tools is that they not only provide the kind of results that Anat describes, but that they often draw attention to their own limitations. Even when black-boxed versions of tools are available, I often encourage students to try to create their own.

Federico Nanni: I would like to further remark on the previous reference concerning tools as black-boxes. On the one hand, computational methods such as LDA topic models are based on specific assumptions, for instance, the fact that documents are generated by a distribution over latent topics, where each topic is characterised by a distribution over all the words in the corpus (cf. Graham, Weingart, & Milligan, 2012). This is the "idea" of what a topic "is", upon which the algorithm is built. If we disagree with such definition, we should not employ LDA in our study to capture "topics". On the other hand, it is true that in the last 10 years LDA has been used many times for explorative/serendipitous analyses, but I would argue that other techniques, which are often simpler to use, faster to run and rely on simpler assumptions could also be very useful, starting from key-phrase extraction (Moretti, Sprugnoli, & Tonelli, 2015).[3] So, in my opinion, researchers should not employ a computational tool because it is widely adopted in the community, on the opposite they should critically question it, especially because it is so widely used.

William J. Turkel: Strongly agree with Federico on this point. I might even say that the more widely adopted a tool is within a community, the less useful it becomes. This is another case where learning new computational techniques can lead to new questions. If one searches through the literature for alternatives to, say, LDA topic models, one not only learns new computational techniques but begins to see the advantages and disadvantages to various approaches.

Niels Brügger and Ian Milligan: If we may quickly interject here – the "more widely adopted a tool is within a community, the less useful it becomes" point that Turkel and Nanni are making is we think worth pausing on. Some work has been trying to coalesce around a standard stack of text analysis tools (i.e. let's make sure all scholars can do X, Y and Z), but now you're making us wonder if this is misguided?

William J.Turkel: When we come up with lists of things that all scholars should be able to do, the key thing is to focus on giving them skills to create tools, rather than giving them tools per se. To use an analogy from basic statistics, the mean is a useful

operation for some datasets but it is not robust because it is sensitive to outliers. If we give everyone in the community a calculator with a button to compute the mean, we soon end up with a situation where the tool gets misused from time-to-time. It is a black box. A better tool would compute the mean while providing the user with some indication of whether it was appropriate or not. But best of all would be a standard of training that allowed people to build and test their own tools, or to verify that the tools of others were working properly and appropriate to the task at hand.

Federico Nanni: I completely agree with William. I recently organised a workshop on tool criticism, where we wrote a dictionary-based approach for sentiment analysis and plot trajectory from scratch (similar to Syuzhet) (Jockers, 2017). By doing so, the participants noticed how many little assumptions are already embedded in just a few lines of code and how by simply changing them, you could drastically modify the final results.

Richard Rogers: There is a series of contemporary critiques associated with the computational approach to the study of digital media, especially given the Cambridge Analytica scandal and the fake news debacle. In the one, a big data approach to studying people's preferences from social media data was unethically repurposed into a right-wing amplification project targeted at keyword publics, and in the other, platforms for participatory culture were reused for influence campaigning and astroturfing. Both of these campaign projects benefited from the knowledge of medium specificity as well as "web- and platform-native" techniques, and have led increasingly to "locked platforms" with social media data becoming scarcer rather than "bigger". There are now calls for "post-API" research, which in a sense is also an invitation to return to small data, ethnographic, and interface methods, including scraping. But web and social media companies actively work against data collection techniques (as well as apps) that they have not approved. For example, researchers who scrape are treated like any other "spammy" actor, and also could compromise themselves by becoming banned or suspended when striving to make a more robust dataset. Instead of allowing data collection through the APIs, social media companies are now furnishing datasets (e.g. in the case of Facebook's Social Science One initiative or in Twitter's sets of Russian and Iranian trolls) but these are "company-curated" and may be critiqued as such. This company-driven research also may lead to particular types of data and analytical practices, e.g. in the Social Science One initiative at one's disposal are all web URLs that have been posted on Facebook in the past year; these may be analysed remotely only on Facebook's infrastructure, with aggregates as outputs. These are big data for computational techniques. Qualitative and mixed methods researchers are not necessarily the envisaged users.

William J. Turkel: On this point, see Eriksson, Vonderau, Snickars, and Fleischer (2019). The authors explicitly engaged in covert and experimental methods (that violated the company's Terms of Use) to explore the "back end" of the streaming music service Spotify.

Matthew Weber: The point about black boxing of the methods cannot be understated. It is important to understand what the inherent assumptions of a methodology

are but also to understand black box issues associated with the data. For instance, I frequently work with social network analysis, and I often see scholars presenting social network diagrams of large datasets without thinking about the algorithm used to generate the diagram. For instance, the popular Fruchterman Reingold algorithm is a force-directed layout that can be quite useful for visualisation, but with large datasets, it struggles to capture differences within subclusters in the data. Always be aware of the limitations and restrictions of an approach. With regard to data, I often work with archived internet data. Archived internet data provide a rich and robust record of web activity and web page content. On the other hand, archived internet data rarely provide accurate summaries of how the data were collected, and what the scoping and limitations are associated with the data. In one recent study, I conducted looking at archived internet data mapping US Congressional webpages, we found that up to 60% of the webpages linked to by Congressional webpages were not present in the dataset. This is not an uncommon problem, but it means that when you are using computational methods in the context of internet data you need to know the limitations of your data. The same is true when you are using pre-existing data such as Twitter and Facebook data, or if you are collecting and scraping your own data. These limitations are fatal flaws; they are often to be expected. But researchers have an obligation to be clear and transparent about these limitations and to provide access to their data where possible so that others can replicate and validate their work.

Sophie Gebeil: I fully agree with the critical and transparent perspective that the social scientist must build on the computational tools they use and the data on which these tools are applied.

On the issue of the "democratisation" of certain tools in communities that could undermine their usefulness, I identify two levels. The first is the need to create tools adapted to new research questions that will inevitably be specific and will also allow innovative methods and results. But at a second level, researchers who have a good knowledge of the tools and their limitations can also help to improve the development of internet studies among beginner researchers in computational methods. To use the case of web archives, I am currently working with engineers from the National audiovisual institute (Ina) on the WebTV collection. The idea is to design an extraction tool to explore this fragmented video corpus according to my questions about the memories of social movements. But there is also the will, or even the requirement for Ina, to propose a standardised tool, which any beginner researcher can use to explore corpora in the Ina archives, while being mindful of the limitations of the tool and these data.

Niels Brügger and Ian Milligan: To use computational methods, the object of study needs to be in digital form. Do you have any thoughts about to what extent the process of collecting influences computational research? Are the right sources collected? In the right format? By the right institutions?

Sophie Gebeil: In case of the web archiving process in France, two Institutions (the Ina, National audiovisual institute and the BnF, French national library) collect the web

within the legal framework set by legal deposit. In my opinion, this is an opportunity because it gives a solid framework for collection institutions and gives web sources the status of a common heritage even if this results in a territorialisation of the web. The Ina and the BnF are obliged to communicate on how data are collected, structured and stored. Moreover, as Valérie Schafer and Francesca Musiani have argued, several pieces of information are missing from the collection process (Musiani, Paloque-Bergès, Schafer, & Thierry, 2019). In return, this also makes it more difficult for computer scientists to use the data, who must necessarily be part of a state-funded research project. Finally, the question of formats is crucial. In France, the BnF uses the WARC format and the Ina has its own internally developed DAFF format. This means that a project to develop computational methods for French web archives should adapt to these two distinct formats, without necessarily being able to cross-reference the results.

Anat Ben-David: I think that several years after what has been termed "the computational turn" in Digital Humanities, or the hype around "Computational Social Sciences", there is already wide acknowledgement among researchers that data collection practices are never neutral, and that constraints on access and on the ability to use various data pose significant challenges to the types of research that can be done with them.

Federico Nanni: The use of computational methods for internet research and the study of our present times is tightly interconnected with the availability of big data to be analysed by the community, from collections of news articles to tweet corpora up to national web archives. An aspect that is, however, not very often discussed is the complexity of obtaining access to such collections, especially for a scholar who is not affiliated with a national library or directly involved in an international project on the topic.

In my doctoral research, I examined the difficulties of reconstructing the history of the University of Bologna website (Nanni, 2017), which was excluded from the Internet Archive's Wayback Machine and at the same time was not archived by the Italian Central Library, as Italy does not systematically archive its national web sphere. The same issue can emerge again when the intention is to retrieve data from social media platforms (e.g. all tweets posted during the first week of Occupy Wall Street) or when we aim to study a specific sub-collection of a web archive (e.g. all personal blogs preserved by the UK Web Archive mentioning the Brexit referendum); in many cases, obtaining these data is at the same time not straightforward and not cheap.[4] This is due to many different reasons, from privacy and copyright constraints to the too often underestimated computational difficulties in retrieving such materials, up to economic interests of the "data-owner", especially when this is not a public institution.

Richard Rogers: In recent years, web data have become "cleaner" in the sense of being pre-structured and well-formatted. Web data are now a far cry from the messiness associated with incomplete crawling, folksonomic labeling and multifarious styles of engine querying. But the "editing" that has resulted in the new cleanliness is also different from how human editors classified websites to make web directories or

Wikipedians removed vandalism or handled troll contributions. From the Wayback Machine of the Internet Archive and Wikipedia to social media companies, data are preferably delivered through APIs, meaning there are available fields in the database as well as query routines. The APIs are designed with particular use cases (or "business cases") in mind, but also are interfaces to back-ends when researchers may have been more familiar with front-ends. When one begins building software that explores the API or wishes to make use of its data for specific research purposes, one also enters the realm of computational and developer culture.

William J. Turkel: On the one hand, there are the institutions like the ones that Richard mentioned which provide APIs and access to masses of downloadable, structured data. On the other hand, it continually becomes easier for individual researchers or small teams to crawl and index significant portions of the web by themselves if they have scripting skills. The datasets that one obtains with the latter method tend not to be clean, and intellectual property considerations preclude sharing them. For one of my research projects, I collected on the order of a few million documents relating to the history of electronics from the open web. The crawling took a few months using nothing more than a laptop and external hard drive. I would not share the dataset publicly, however, because I have no metadata associating each item with its copyright status.

Matthew Weber: Social media are a bit more straightforward when it comes to the sources and format, as well as the institutions. The structure of the content leads to a somewhat common pattern of data storage; for instance, most Twitter data contain certain key data points about the user and the content of tweets. Data today are often cleaner in terms of the formats; we have better standards for data structure than we did a decade ago, and we have a better awareness of those standards. At the same time, there are so many different types of data, and different standards, that the problem of data format remains complex. Web archiving is a great example; the web archive (WARC) file format is a standard for storing archived web pages, and yet different institutions use the WARC in different ways (populating some fields, and not others, and specifying provenance in a variety of ways). Data shared by the Internet Archive often differ from data shared by the Library of Congress or the BnF, and access varies widely as well. To Federico's point, access is often an issue as well. Twitter is a great example of access issues; the degree of access (and the percent of data you are able to view) varies widely based on access to funding, access to the data firehose (or API), and existing relationships with researchers at Twitter.

Niels Brügger and Ian Milligan: Is there anything that in your mind impedes the use of computational methods in studies of the history of the internet? Are source collections not "researcher-friendly"? Is there a lack of adequate methods and tools? Are there other obstacles?

Sophie Gebeil: As I have worked with web archives since 2011, I am finding that the modes of access have diversified. Of course, there is still a lack of tools that would allow us to quickly identify some characteristics from a corpus: hyperlink dynamics, textual analyses or even image analysis. It is difficult because of the specificities of the

web archives but it is a new and exciting field. In my opinion, it is also the role of researchers in human and social sciences to contribute to the design of this type of tool. There are more and more historians who are interested in digital sources or digital humanities. There are also reticences that come primarily from a lack of training in computational methods, but another reason is also the very strong relationship that the historian has with the document, the archive. These two notions are disrupted with born and reborn digital sources.

Anat Ben-David: Reading the literature on web archives, one comes across the word "challenges" very often. Not only is the web a medium that is challenging to preserve, the solutions that have hitherto been proposed to archive it result in collections that do not lend themselves easily to computational research. As I previously mentioned, since the internet is a computational medium to begin with, it is rather tempting to try to apply computational analyses for studying its history.

But one finds out very quickly that even the simple computational methods that are used for studying the live web, or other digital data, cannot be applied to the archived web, for the following reasons. First, copyright and privacy constraints limit access to web archives, which is often restricted to viewing at the premises of national libraries. Second, current interfaces to web archives are primarily designed for viewing, or "surfing" archived snapshots (one page at a time), but not for their treatment as corpora. Third, archived web materials often lack sufficient context on the circumstances and techniques of the archiving; without such provenance information, it is difficult to determine which snapshots should be included or excluded from the analysis, or to explain temporal incoherence or "holes" in web archives. Finally, most infrastructures that are currently being used to host very large web archives are not designed to support computational processing.

Federico Nanni: I agree with the point made by Sophie on the fact that humanities/social scientists should contribute to the design of tools. However, it is often not easy for a humanities researcher to develop the computational skills and data science knowledge actually required for contributing and therefore moving out of a setting where they are simply "computer scientist customers", as Adam Crymble once put it (Crymble, 2015). Quite often Digital Humanities researchers spend their entire doctoral studies in building up such expertise, and they might never have the time/chance of actually using such tools in a substantive research.

William J. Turkel: But Crymble is an excellent example of someone who started developing computational skills as a graduate student, and continues to do so, and to share what he has learned.

Federico Nanni: As I mentioned before and as Anat has also remarked on, I believe that the number one issue is the prompt accessibility to web archive data. While this is due to understandable reasons, the lack of access for the broad academic community has limited, among other things, the development of tools specifically tailored for particular web archiving issues and, I would argue, also the perception of the challenges that web archivists are currently facing. For instance, information retrieval approaches that address the complexity of dealing with different temporal-layers of a

web archive are often developed by research groups having direct access to a web archive (for instance through a research project or a collaboration with a web archive).

A tool that, at least in my opinion, seems to be highly needed by both web archivists and internet studies researchers is a method for building topic-specific collections from a web archive (e.g. by isolating only pages referring to a specific event); this would, therefore, produce a smaller and hopefully manageable sub-collection that the researcher could further study on its own. However such complex methodological challenge is currently not fully addressed by the broad information retrieval research community.[5]

Richard Rogers: Most histories of the internet have been written without computational techniques, just as most histories generally. Thus, the internet is not special in that regard. Moreover, many histories of the internet have been written without (citing) web archives, which could be considered a main source of historical material. These archives also could be the site for computational techniques and tools. To date, however, the computational study of web archives has been quite distinctive from internet and web historiography. This state of affairs may be changing, as there are examples of collaborations between digital methods researchers and internet historians as well as between web archivists and internet historians, though projects with the three parties could be developed.

Matthew Weber: One question that comes to mind is whether there is anything unique or remarkable about our ability to "replay" the internet? Is it enough to replay an image of a webpage, or do you need access to the underlying code? In other words, do you need to see the technology as it was constructed? I would argue that for the internet the code and technology are critically important.

William J. Turkel: A very interesting question. Since web page Mementos are reconstructed from components that were archived at various times (or not), the "fidelity" of the playback is limited and the reconstructed pages are at best an approximation of the past. That said, they are a valuable and widely accessible resource.

Matthew Weber: I believe there are many collections that are researcher-friendly. The Internet Archive's Wayback Machine is a wonderful interface for viewing the history of a webpage from a qualitative point of view. Their research services team has developed tools designed to improve access to a subset of data. In the academic space, the Web Science and Digital Libraries Research Group at Old Dominion has produced a great collection of computer science oriented tools for research and access. The Archives Unleashed team at the University of Waterloo has also developed tools that are much easier to use than what we have seen in prior years. At George Washington University, Social Feed Manager allows researchers to work with social media data, and to collect their own datasets. I believe we have a host of tools from different disciplines that have enabled access and opened up researcher access to large-scale data. These tools, however, often require technical skills in order to wrangle the tool to return the desired results. As is common with this type of computational work, I believe that interdisciplinary collaboration is a key to success.

William J. Turkel: I once read somewhere that the most successful interdisciplinary work happens when a single individual is trained in the techniques of multiple disciplines. While I agree with Matthew's point that interdisciplinary collaboration is a key to success in computational work, I also think that each of us needs to strive to be "a kind of import-export specialist between the disciplines" as James Clifford put it (Clifford, 2003, p. 55). More than once, I have seen historians or other humanists expect their technical collaborators to get up to speed on the literature of the topic, while resisting any engagement on their own part with the literature of the technical methods. Those kind of collaborations tend to fail.

Peter Webster: To reiterate a point made by other contributors, for me the biggest single challenge (of the many) is the fragmentation of a medium that only very loosely behaves in a "national" way into nationally conceived archives. As I've argued elsewhere, the pattern into which web archiving has fallen has its own history, and we have to be very thankful for non-print legal deposit since without it we would be even more reliant on the Internet Archive than we already are. But in order for us to effectively study the Web in a way that aligns with its fundamental nature, we need methods of transnational discovery and analysis, and if that necessitates government-level action to amend copyright legislation in different nations, then we should be lobbying for that. (We perhaps stand the best chance of success at the level of the European Union.)

Niels Brügger and Ian Milligan: How do you see the relation between subject-matter experts like historians and new media scholars and developers (from systems librarians to programmers)? Should internet historians learn to code, or conversely, is the onus on developers to learn about historical methods?

Sophie Gebeil: It depends on the research field. From a cultural history perspective, it does not seem essential to me to know how to code but it is necessary to have some knowledge of coding and HTML language in general. If the "programmer historian" (Le Roy Ladurie, 1973) exists, in my opinion, they will remain a minority. Above all, I believe that historians must develop a digital culture and computer skills in order to be the best possible interlocutors to participate in the design of computational analysis methods with developers. In return, I find it interesting to consider the fact that computer scientists also develop a culture in the human and social sciences even if I had never asked myself the question in this sense. In my research experience, I had to learn to use Navicrawler, Hyphe and Gephi on my own and therefore I use them in an approximate way. It seems to me that one of the challenges is precisely to succeed in cooperating all together (historians, engineers, archivists, programmers, etc.) to propose innovative methods but also easily usable tools that would democratise the use of natively digital sources in history.

Anat Ben-David: From my experience working with developers, information scientists and computer scientists, the interdisciplinary collaboration is successful when the research questions, or the object of study, are interesting enough – scientifically – to all involved. Computer scientists might not be interested in a historical question if they do not find the computational challenge interesting enough. Reversely, historians do not necessarily need to learn to code to conduct simple computational analyses,

but they may benefit from knowing the types of available analyses that can be performed, and how to communicate to developers what they would like to achieve with computational analyses.

Federico Nanni: I agree with the opinions above, but I tend to have an even more pragmatic view on the topic, probably influenced by the fact that I have spent the last five years as the only (digital) historian in a data science research group. It is true that the research question needs to be perceived as "interesting" by the computer scientist and the computational aspect of the problem needs to be "challenging enough", but I think that this is often not the main issue.

The problem and the developed approach also need to be in line with the methodological interests of the computer science community of reference (in my case often the natural language processing community). This means that if an interesting and challenging problem could not – for instance – be successfully addressed with a deep learning architecture (e.g. due to the sparsity/lack of training data), but instead with a more traditional word frequency-based approach, such work will suddenly become way harder to be published in a Natural Language Publishing (NLP) venue. Subject-matter experts and computer scientists starting a collaboration should not underestimate the complexity of such settings, especially for what concerns the publication process; the two communities have very different practices, from pre-print publications to data-sharing up to established policies concerning author names on research articles.

At the beginning of my Ph.D., I was very much guided by the idea of becoming a "programming historian" able to conduct my research in a completely independent way, benefiting from the knowledge of the two communities. However, developing a proper data science profile is actually very challenging and it could bring you far away from the research question that you originally intended to address, often to a place where it is difficult to demonstrate the relevance of your research to either community, because it is at the same time not "novel" enough for an NLP audience and not "substantive" enough for a historical one. I still believe that we need a generation of programming (internet) historians, for critically addressing many of the new challenges of dealing with web archives while at the same time pursuing historical scholarships, but we especially need very well prepared interdisciplinary educators.

Richard Rogers: I am reminded of the critique of the computational turn in humanities that invites those learning a corpus to at once be trained in analytical software operation. As a rejoinder, "button-pressing" is defended in the history of the humanities these days as contiguous with the long tradition of pattern recognition that has developed alongside hermeneutics. Thus, there always have been humanities coders, albeit in relatively smaller numbers. Perhaps the question concerns how central computational culture should be (and how strident the response) in the development of curricula and the larger programmatic agenda, which these days is favouring such work in new funding, career and other schemes.

Matthew Weber: Anat and Federico both raise important points of the nature of interdisciplinary collaboration. And as Richard points out, there are always those within

a discipline who are able to translate work across disciplinary boundaries – for example, understanding code such that the barriers to collaborating with programmers are minimised. In my experience, the research questions need to be compelling to all involved in a project, but what is ultimately compelling to one person will not be the same for all others. On a recent project examining the change in local news ecosystems, I was interested in the dynamic nature of interconnections between websites in the data my team collected. The computer scientists were glad to help with the problem of coding the way I tracked these evolving networks but found the computational work to be relatively simplistic. Rather, their interests were driven by a desire to use the corpus of text as a way to trace patterns of misinformation through a network of news. The dataset and project were both sufficiently large as to allow each member of the team to carve out his or her niche. In sum, I believe it is important that each team member find their own motivation. I don't expect that historians will automatically learn code, or that computer scientists will learn the nuances of digital humanities scholarship, but it is important to find a common language. Understanding in both directions will ultimately increase the success of the research.

William J. Turkel: Looking at the question slightly differently, I would argue that *all* undergraduates should be encouraged to try programming to see if they like it. Assuming that the vast majority of our students will never become practicing historians, it is great to have skills that pay the rent. Whenever someone asks me for career advice I encourage them to develop marketable skills that resonate with them. Learn another natural language; learn to program; take a course on linear algebra, statistics, accounting or finance; learn GIS or databases, etc.

Peter Webster: I think there are two distinct questions in play here, although they are related. The first concerns how research projects are conceived in terms of their staffing, which in turns depends on models of funding. If a project is led by historians, and it is in history or other humanities disciplines that the research interventions are to be made, then the relationship with developers will most likely be one of contractor and client. If the scholar is able to articulate their requirements clearly enough (though this is very often not the case), then the relationship is relatively easy to manage. If, on the other hand, the project is conceived as one which speaks to both questions in the humanities and in computer science or library and information studies, then the dynamics will necessarily be different, as in the cases that Federico and Matthew outline. So, whether the onus is on the development side depends on the kind of project.

More generally, the question "should historians learn to program" is a slightly unhelpful one. If we were instead to ask: "do we need there always to be *some* historians who are learning to program", then the answer is clearly a positive one. And this, as Richard rightly observes, has been the case for many years, as small communities of scholars find and experiment with new tools and approaches as they appear, and show to the rest of the discipline what might be possible. Those people will continue to select themselves by the route that William describes. But it is (I think) neither possible or desirable for all historians to be proficient programmers, since the diversity of what it means to be a developer is already very great, and likely to become greater.

Even before one generation of scholars has mastered one language (or even begun to get to grips with it) it will have changed, and alternatives sprung up to replace it. What scholars do however need, I think, is a grasp of basic principles of computer science, data management, archival science, project management and (in particular) of the characteristics of successful development projects.

Niels Brügger and Ian Milligan: The use of computational methods in historical study has a history of its own. What are the most defining moments in the history of computational methods?

Anat Ben-David: That's a tough question, especially considering that media (and science) histories are not necessarily linear, or defined by key moments. But if I must answer, then I would note the development of cloud computing that allowed scaling analyses beyond the constraints of physical memory, and the development of open source programming languages such as Python and R, that attracted a wide community of users.

Federico Nanni: It's always a matter of considering computational as a sort of synonym for quantitative methods or not, which could open a never-ending digital humanities discussion. Apart from the most famous turning points in the relations between the quantitative/computational and the historical (from the discussion around "Time on the Cross" to the Google Culturomics project), I consider defining moments for our discipline all the improvements in information retrieval systems, and their impact on our everyday life and our work as historians. From Karen Sparck Jones' inverse document frequency term weighting to PageRank up to the more recent "things not string", these are all technical innovations that have influenced (and often improved) our digital archival research and consequently our scholarships.

Richard Rogers: The debates these days around digital humanities remind me of controversies surrounding the introduction of cliometrics in the 1960s. Cliometrics put a name to the use of quantitative methods in historiography, applied especially in economic history (and also the history of technology). To me, the interesting aspect of cliometrics was less the introduction of stats and data to history and the debates around disruptions to fields and paradigms, but rather the style of the research questions. The more well-known work employed counterfactual historiography. What if the American civil war did not take place? What if the railways were not built? The latter question concerned whether the canal system could have led to similar levels of economic development as the railways and the second industrial revolution in the USA. The point was to question the "axiom of indispensability" of the railways to development. (Robert Fogel won the Nobel Prize for Economics with this and other work.) As it matured cliometrics was no longer a movement in "new economic history" and rather experienced a typical pioneer's regress, becoming again a branch rather than a trunk route in the larger field of historiography.

No one is asking whether digital humanities would suffer the same fate or enjoy the same prizes, but the introduction of the quantitative, statistical, computational and similar instrumentaria, together with their styles of inquiry, could lead to temporary novelty or pockets of innovation.

Matthew Weber: The continued growth of a robust community of researchers working with Python and R has been a key development for computational methods. The community of scholars working in this space is generous with their time, and work to share best practices and code. I think there are other technologies, as well, that are helping to lower barriers. Interactive what you see is what you get platforms, such as Python notebooks, allow you to see code and output together in a seamless interface. Beyond that, there has been a groundswell in workshops and tutorials at annual meetings, over the summer, and online, that has served to create a rich set of educational resources.

Niels Brügger and Ian Milligan: How do you see the future of using computational methods for historical studies of the internet? What are the biggest challenges? The biggest opportunities or most exciting projects today? Which type of methods and tools would you like to see developed?

Sophie Gebeil: There are several perspectives to which I am committed.

On the methods and tools side, I am currently working with Ina on data extraction from web archives. This is important to me because I believe that the creation of corpus analysis tools would facilitate the appropriation of web archives by researchers in the social sciences and humanities. In the future, I would like to see the development of methods related to visual studies that allow the identification of the path of visual content from pre-existing media archives (print media, television) to and in the archives of the Web. For example, we could then trace the video or fixed image of General De Gaulle pronouncing the famous "I understood you" on 4 June 1958 in Algiers and follow the circulation of the image, its diversion in the web archives of the BnF or the Ina, or in other web archives.

Another aspect that is close to me is the development of a reflection on the archiving of the Web in the Mediterranean, which to date is mainly the work of scattered and isolated groups. In a context of instability and major political changes, the collection, preservation and study of natively digital sources is a fundamental challenge for Mediterranean societies.

Anat Ben-David: As Niels Brügger noted, web archives are not exactly archives, since their organisation and structure lack archival principles such as appraisal and provenance. In that sense, future computational methods may be helpful in improving archival appraisal and in adding provenance and other contextual information that may significantly increase the utility of the archived web for historical research. Important computational work is currently being conducted by the Memento project at Old Dominion University and elsewhere, where researchers develop methods, tools and web services for understanding the archived web beyond the boundaries of a single collection or archiving institution; and by "The Archives Unleashed" project, led by researchers from the University of Waterloo, which develops toolkits that facilitate the analysis of large scale web archives for historical research. There are two areas that require developing new methods and tools: the first is the question of web archiving after social media, and how to facilitate research across different types of web archives and other datasets, and the second is the need to develop tools specifically designed for critiquing the archiving process, or web archives as institutions.

Federico Nanni: As I remarked before, I believe the core challenges for the future of computational methods in historical studies are twofold: on the one hand, the difficulties of accessing (and therefore experiencing) web archived collections and on the other hand, the lack of critical attitudes towards computational methods. For these reasons, hands-on activities such as the ones organised by the Big UK Domain Data for the Arts and Humanities (BUDDAH) project and the more recent Archives Unleashed series are absolutely essential. I took part in the first two editions of the Archives Unleashed and that was an incredibly formative experience, especially because I had the opportunity of facing for the first time many of the issues that I knew only from literature. Another challenge that I believe is necessary to address involves extending the use of computational methods and web archives to other disciplines, first of all to researchers in political science, which would largely benefit from obtaining a novel diachronic perspective on party politics, international relations and overall democratic processes.

Richard Rogers: National libraries, perhaps understandably, are treating the world wide web as a national web and archiving only their "home" webs. Social media platforms are not providing public archives and are not being publicly archived. The encrypted ephemerality of messaging apps provides another challenge. These days it's as if much of the content, however valuable, is out of reach of the archivist. I'm buoyed by the increased usage of web archives by scholars and students and would encourage developing and also compiling teaching units with web archives.

Matthew Weber: Continued advancements with regard to research at scale will allow for new questions to be asked. We do not fully know the scale of the internet and the web, because we have not yet been able to crawl and analyse the full extent of the web. In this way, we do not know the full extent of the web, nor the history of many aspects of society and interaction on the web. Simultaneously, the tools that allow us to navigate through collections and to extract subsets continue to be developed. This is a burgeoning area of research and as more scholars come into this space it is clear work will expand into new domains. As noted by Anat and Federico, I believe that the current push for educational resources is the fields greatest strength, as new scholars will continue to push the domains of computational methods and internet research.

William J. Turkel: For me, one of the more interesting research areas right now has to do with the adaptation and development of sublinear algorithms that allow us to analyse internet phenomena at scale and/or in real time. One excellent recent example of this kind of work is Ben Schmidt's analysis of the approximately 13.6 M books in the Hathi Trust collection using a general-purpose dimensionality reduction that is ultimately based on the Johnson-Lindenstrauss lemma (Achlioptas, 2003; Schmidt, 2018). Another interesting research area is the development of increasingly high-level languages that encapsulate tens or hundreds of thousands of pages of low-level code into "superfunctions". The Wolfram Language (aka *Mathematica*), for example, allows programmers to implement supervised and unsupervised machine learners with a line or two of code (Bernard, 2017).

Niels Brügger and Ian Milligan: We're seeing a good note of optimism here, as we talk about how tools and programming languages are improving, new research questions can be asked, albeit with some challenges. As this new field comes together, we wonder if we might close our round-doc by asking if you had any recommendations or thoughts for scholars entering this new field? Beyond whether they should learn to program or not, what advice would you give a new entrant to the field?

Matthew Weber: I believe most would agree that computational methods applied in research contexts related to the history of the internet, to digital humanities, or the social sciences, has the potential to open up new avenues of research. I'm optimistic that we are working in an era of academic innovation, and that as scholars working with computational methods we have the opportunity, on the one hand, to look at existing questions in new ways, and, on the other hand, to ask new questions and build new theory. In my current work on local news, my team has been able to look at large subsets of a media ecosystem than ever before, allowing us to analyse and theorise about broad patterns of change with a greater degree of accuracy. Prior work in this space has generally been limited to a single cross-section or a subset of media. While I am knee deep in data, I am energised by the way in which we are able to take a fresh look at questions scholars have been grappling with for decades.

William J. Turkel: I couldn't put it better than Jonas Salk (who was paraphrasing Socrates): "Do what makes your heart leap rather than simply follow some style or fashion" (Salk, 1991).

Federico Nanni: Get in touch with the research community as early as possible, by going to a conference (and RESAW might be the perfect choice) or taking part in an Archives Unleashed Datathon! For me, both have been incredibly enriching experiences during my Ph.D. research.

Sophie Gebeil: The first thing I will say is "you are not alone", there are dynamic research communities on internet studies, the history of the Web in relation to computational methods that are at the origin of an important historiography. Second, I think that neophytes also need to trust each other because innovation comes from taking risks, meeting people, but also sometimes from questions that seemed candid.

Anat Ben-David: It is exciting to learn new methods for historical research. But following William's quote, what "makes my heart leap" is the old-fashioned excitement of archival discovery, and finding the common thread between the computational analysis and the historical narrative.

Notes

1. https://web.archive.org/details/histoire-immigration.fr, consulted on 29 March 2019.
2. See, for instance, how Jockers (2011) described Macroanalysis.
3. https://dh.fbk.eu/technologies/kd
4. See, for instance, the pricing for using the Premium Twitter API, which gives you the possibility of searching the full archive: https://developer.twitter.com/en/pricing.html
5. This is because web archives are a very different type of collection compared to for instance a newspaper archive, upon which traditional topical filtering algorithms are usually tested by the information retrieval community.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Achlioptas, D. (2003). Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, *66*(4), 671–687. doi:10.1016/S0022-0000(03)00025-4

Ben-David, A., Amram, A., & Bekkerman, R. (2018). The colors of the national Web: Visual data analysis of the historical Yugoslav Web domain. *International Journal on Digital Libraries*, *19*(1), 95–106. doi:10.1007/s00799-016-0202-6

Bernard, E. (2017). Building the automated data scientist: The new classify and predict. Wolfram Blog (10 October 2017). Retrieved from https://blog.wolfram.com/2017/10/10/building-the-automated-data-scientist-the-new-classify-and-predict/

Clifford, J. (2003). *On the edges of anthropology: Interviews*. Chicago, IL: Prickly Paradigm Press.

Crymble, A. (2015). Historians are becoming computer science customers – Postscript. Digital History Seminar. Retrieved from https://ihrdighist.blogs.sas.ac.uk/2015/06/historians-are-becoming-computer-science-customers-postscript/

Eriksson, M., Vonderau, P., Snickars, P., & Fleischer, R. (2019). *Spotify teardown*. Cambridge MA: MIT Press.

Graham, S., Weingart, S., & Milligan, I. (2012). Getting started with topic modeling and MALLET. The Programming Historian. Retrieved from https://programminghistorian.org/en/lessons/topic-modeling-and-mallet

Jockers, M. (2011). On distant reading and macroanalysis. Retrieved from http://www.matthew-jockers.net/2011/07/01/on-distant-reading-and-macroanalysis/

Jockers, M. (2017). Introduction to the Syuzhet package. Retrieved from https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html

Le Roy Ladurie, E. (1973). L'historien et l'ordinateur. Le Nouvel Observateur, 05/08/1968.

Lyotard, J.-F., & Chaput, T. (1985). Épreuves d'écriture. Paris: Éditions du Centre Georges Pompidou. Retrieved from https://monoskop.org/images/f/f9/Les_Immateriaux_Epreuves_d_ecriture.pdf

Moretti, G., Sprugnoli, R., & Tonelli, S. (2015). Digging in the dirt: Extracting key phrases from texts with KD. Proceedings of the 2nd Italian Conference on Computational Linguistics, 3-4 December 2015, Trento.

Musiani, F., Paloque-Bergès, C., Schafer, V., & Thierry, B.G. (2019). *Qu'est-ce qu'une archive du web?* Marseille: OpenEdition Press.

Nanni, F. (2017). Reconstructing a website's lost past: Methodological issues concerning the history of www.unibo.it. *Digital Humanities Quarterly*, *11*(2), 1.

Salk, J. (1991). Interview with Jonas Salk. Academy of Achievement. Retrieved from https://www.achievement.org/achiever/jonas-salk-m-d/#interview

Schmidt, B. (2018, October 3). Stable random projection: Lightweight, general-purpose dimensionality reduction for digitized libraries. *Journal of Cultural Analytics*. doi:10.31235/osf.io/36neu