International Journal of Digital Humanities (2019) 1:251–262
https://doi.org/10.1007/s42803-019-00014-8

RESEARCH ARTICLE

Editing social media: the case of online book discussion

Peter Boot[1]

Published online: 22 May 2019

Abstract
Online book discussion is a popular activity on weblogs, specialized book discussion sites, booksellers' sites and elsewhere. These discussions are important for research into literary reception and should be made and kept accessible for researchers. This article asks what an archive of online book discussion should and could look like, and how we could describe such an archive in terms of some of the central concepts of textual scholarship: work, document, text, transcription and variant. What could an approach along the lines of textual scholarship mean for such a collection? If such a collection holds many pieces of information that would not usually be considered text (such as demographic information about contributors), could we still call such a collection an edition, and could we call editing the activity of preparing such a collection?The article introduces some of the relevant (Dutch-language) sites, and summarizes their properties (among others: they are dynamic and vulnerable, they contain structured data and are very large) from the perspective of creating a research collection. It discusses the interpretation of some essential terms of textual studies in this context, and briefly lists a number of components that a digital edition of these sites might or should contain. It argues that such a collection is the result of scholarly work and should not be considered as 'just' a web archive.

Keywords Online book discussion · Edition · Textual scholarship · Born-digital text · Web archives

1 Introduction

Online book discussion is a popular activity on many social media sites: on weblogs, in Facebook groups, in forums, on specialized book discussion sites such as Goodreads and,

✉ Peter Boot
peter.boot@huygens.knaw.nl

[1] Huygens Institute for the History of the Netherlands, PO Box 10855, 1001 EWAmsterdam, The Netherlands

perhaps most voluminously, on booksellers' sites, such as Amazon. These discussions are important for research into literary reception (Boot 2013) and, therefore, should be made and kept accessible for researchers. In this article, I ask what an archive of online book discussion should and could look like, and how we could describe such an archive in terms of some of the central concepts of textual scholarship: work, document, text, transcription and variant. The discussion is based on experiences gained in the creation of an archive of Dutch online book discussions (Boot 2017), a collection that now holds 870,000 items of book response downloaded from a number of different sites into a single database. What could an approach along the lines of textual scholarship mean for such a collection? Conversely, perhaps, what could a study of these sites mean for textual scholarship? If such a collection holds many pieces of information that would not usually be considered text (such as demographic information about contributors), could we still call such a collection an edition, and could we call editing the activity of preparing such a collection? Or is 'thematic research collection' (Palmer 2004) or 'archive' (Vanhoutte 1999) a better name? Does it matter which name we use (Price 2009)?

I will first introduce some of the relevant sites, and I will summarize their properties from the perspective of creating a research collection. I will then give an interpretation of some essential terms of textual studies in this context, and I will briefly list a number of components that a digital edition of these sites might or should contain. This will prepare us for a discussion of the wider issues raised above.

2 Dutch online book discussion sites

At the moment of writing, the most prominent Dutch book discussion platform is review site Hebban (Jessen 2016), comparable to Goodreads in the English-speaking world (Thelwall and Kousha 2016).[1] Hebban claims to have 250,000 unique visitors per month, 139,000 registered users, who have left 817,000 ratings and 90,000 reviews.[2] Readers create a profile for themselves, with demographic data, a short description and perhaps links to other social media accounts. They create book lists, can follow other readers, and can respond to contributions from other people on the site. The site also keeps metadata about books, such as publisher information, a description, ISBN and an aggregate rating. Hebban is not just a collection of texts, but also a large collection of structured data.

A very different way of discussing books is practised on bulletin-board style discussion forums. The largest one in the Netherlands is Ezzulia.[3] The typical structure is an organisation into forums and subforums, divided into topics, which consist of an opening post and subsequent discussion. Here, the interaction between readers is much more important than more or less formal reviews. There are usually no formal metadata about books, but the forum-post-response relation provides the organizing data structure for the forum texts.

The largest online bookseller of the Netherlands is bol.com. It is the only online bookseller where buyers have left a substantial number of user reviews (Daniels 2016).

[1] www.hebban.nl
[2] https://static.hebban.nl/files/files/hebban_prijslijst_2018v20-3LR.pdf. The number of reviews is based on a download of the site in February 2018.
[3] http://www.ezzulia.nl/forum/

We downloaded 250,000 reviews from the site. In terms of facilities for discussion and interaction, the reviews at bol.com have little to offer, and users have no way to identify themselves with a picture or profile text.

The most informal book discussion probably takes place in Facebook book discussion groups. There exist a number of Dutch language book discussion groups, some smaller, some larger. The largest one at the time of writing is Samenlezenisleuker (Reading together is more fun), with ca. 10,000 members.[4] The focus here is on conviviality and sociability, for which books provide the occasion. A prominent characteristic of Facebook groups is that they often span multiple sites: besides their Facebook home, they also have a Twitter presence, an Instagram account and a weblog.

Unlike the sites discussed up to now, the typical weblog is an individual undertaking. Because of that, it is much harder to give a brief characteristic of book blog (a weblog about books) based book discussion. Some book blogs are maintained by teenagers, some focus on genre books (such as suspense novels or historical romances), some are written by people with a lifetime of thinking and reading behind them. Although they are usually maintained by individuals, through their blog rolls (lists to other blogs they are following) and the comment facilities, weblogs and their maintainers are part of a network. This network extends to the other sites we discussed: many bloggers are also active on the other sites, and there are many names that we encounter on many different platforms in the network of online book discussion.

3 Properties of sites

Supposing that we are going to create a research collection of online book discussion material, we should note that the material that we are going to archive or edit is very different from the sort of material that textual scholarship usually has to deal with. There are at least eight ways in which online book discussion is different. First of all, these sites are *dynamic*. Many editorial projects deal with a closed corpus: the author is dead and though we may not yet know all relevant witnesses, no new text can be created. In our case, we deal with dynamic sites that grow each day; new book responses are added, new reviewers arrive, reviews may also be removed. People strike up friendships. An edition of a live site can never be complete. Second, websites are *transient*, they need effort to be maintained. The absence of a material carrier makes web sites very vulnerable, and this lends a particular urgency to their archiving. Some of the larger Dutch book discussion sites have already disappeared. Third, while there are many traditional edition projects that have to handle material in multiple media, for literary editions they are usually not the centre of attention. In contrast, the texts in online book discussion use *rich text*: they use all of the features of HTML, such as hyperlinks, use of other media (images, video, audio), different forms of highlighting and other rendering styles, such as different fonts. An edition of this material will have to know how to handle this from the start. Fourth, these sites are *structured networks* of textual constituents, enhanced by metadata, often a collaboration of many participants. The display of the textual material is determined by the user's interests: reviews, for instance, can often be accessed by book, by reviewer or by tag. Reviews often spurn

[4] https://www.facebook.com/groups/451488498379185/

conversations. Pages are enriched with metadata. As we will see, this calls for an enriched concept of text, which includes the data underlying the navigational possibilities of the sites.

Fifth, sites are *multi-platform constructions*. As we saw for the Facebook group, sites are often parts of a larger network. Beyond that, users often employ multiple sites for different aspects of book discussion: to write reviews, to announce their reviews, to discuss reviews from others, or simply to hang out. A decision to archive or edit a single site is to limit ourselves to a to some extent arbitrary section of the book discussion environment. Sixth, there are *legal complications* in publishing a collection of this material. Even though texts have been published online, copyright has to be respected. In addition, privacy regulations may be problematic. Seventh: depending on the site, the *volume of material* may be very large. Together with the eighth and final property, that collections are *electronic* from the start, this implies that the creation of the collection will have to be automated and that it will be impossible to devote attention to the individual text (although the programs will have to be customized for the different sites or site types). For the user, the importance of the collection will probably be as much in the trends and patterns that the collection shows as in the individual texts, and the collections will have to offer 'distant reading'—like research facilities.

4 In editorial terms

The book discussion sites' properties have consequences for how an edition of these sites should be conceived. I briefly discuss some relevant terminology of textual scholarship in this context, mentioning the relevant properties from the previous section.

Work: the work that we would edit will usually be the site (or, based on the multi-platform property, a site cluster), as it is not the individual reviews that we are interested in but the ways these are organized in structured networks. In the case of mass reviewing sites, this is a highly collaborative work, with some people responsible for the design and maintenance and thousands of people responsible for the content. In terms of the $FRBR_{OO}$ ontology, the site is an 'aggregation work', a work that essentially consists 'the selection and/or arrangement of expressions of one or more other works' (IFLA 2015 65); these other, individual works being the reviews, blogposts, etcetera that the site contains. A weblog, on the contrary, is usually the work of a single person (but can still be considered as an aggregate of its posts). In some situations, it might not be necessary to edit an entire site, e.g. for the reviews on booksellers' sites. It is also clearly impossible to edit Facebook.

State, version: as mentioned, the site, and thus the work, is a dynamic entity. It is therefore also what the $FRBR_{OO}$ ontology calls a 'complex work', a work for which there are multiple versions. A state or version of the work is the site as it exists at any specific moment. It is up to the editor to decide whether the edition should try to document a state of the site, or the development of the site through a given time period. We could call an edition of the former type a synchronic edition, and the latter type a diachronic edition. In any case, for a living site, the edition is necessarily incomplete, unless we devise some way to automatically enhance the edition with the latest live data.

Documents, witnesses: when we ask on the basis of what documents or witnesses we should edit the sites, the answer is not clear-cut. There are three options: the server database, the html and intercepted communication between server and browser, or saved output from API (Application Programming Interface) calls. Each of these options runs into some difficulties because of the sites being transient and dynamic.

In a sense, the ultimate witness is the database from which the server constructs the html pages that the site's users see. That database contains the posts, the user information, the book information, the book lists, etc., but there is a number of reasons why this answer is unsatisfactory. First, the database underlying the site is a live thing, constantly modified, and a witness that we cannot use to check our edition is not very useful witness. A read-only copy might be a better candidate, but there remain other problems. The database, or copies of it, may be out of reach of the editor, in the case where we create a research collection without asking for cooperation from the site. Finally, and perhaps most fundamentally, the site as a publication is defined by the web pages that the user gets to see, and it may be a non-trivial effort to understand how exactly the database content is related to the content of the web pages.

However, when we take our authoritative witness to be what the user sees, we run into other problems. What the user sees— by and large, a rendered HTML page — is sent to the user from the server. The user can save the HTML file to disk to create a basis for the edition, but in the process becomes a co-creator of the witness. In practice, requesting the site's pages and saving them (a process known as 'scraping' a site) will be relegated to a program. But this doesn't alter the point that the witness on the basis of which the edition will be created is no longer input to the edition project, but one of its outcomes. We also have to save any additional files, such as stylesheets, javascript, and images. A further complication here is that a web server may decide to show different things to different users, because of geographical location, hardware characteristics, or many other things. Another complication is that in modern web pages, it is not just HTML that is passed around. Much of the content is fetched using javascript in JSON format. To capture that content, the editor has to write a program to execute the necessary queries against the server and save the outcome, again creating the witnesses that should be the givens in an edition project.

A final option is for the editor to use a published application programming interface (API) to the server collection, if the server offers this possibility. An API offers direct access to the underlying data of the site. It is, in a sense, the cleanest and most direct access to the data that is shown in the site. But similar problems arise as with the previous options: the program accessing the API would have to store the output of the API requests, and this saved output becomes the witness for the edition. That is: the edition creates its own witnesses, and it is not self-evident that content of the published site is identical to the output of the API requests.

In the following, I will assume that saved HTML will provide the basis for our hypothetical research collection. That seems to be the practice in most projects that want to create web corpora. In any case, the witnesses are dated. They document the state of the site at a certain date. If we want our edition to document the development of the site, we'll need to create witnesses at certain intervals, or perhaps set up a continuous monitoring process.

Text: with respect to the text of the edition, we have two terminological possibilities. The first option is to stick to the common-sense, linguistic notion of text, and then the text

of the edition is the collection of its text fragments: the posts, reviews, comments, and so forth, that is: the text of the individual works that the site aggregates. The site as an aggregate work does not have much text of its own, except for things like the help pages, the about pages and the publicity. What defines the site are the structures within which the aggregated text fragments are embedded and the data elements attached to that structure, which include the organisation of the text fragments by book or by reviewer, their dates, and the ratings. In this option, for this type of work, the establishment of a text as a prime task of editorial endeavour is joined on an equal footing with the establishment of the work's structure and data. We should also note that even in this restricted notion of text, it has to be understood as 'rich text' because the text fragments can contain illustrations, lists, hyperlinks and all the other features that HTML offers.

The second way of understanding the word text in this context takes into account the network character of these sites, and sees text as the structured whole of data and text fragments that forms the content of the site. From this point of view, we can safely say that text remains the main object of editing, but we should be aware that this text is in no way identical to the text as a concatenation of all the text fragments. As I take the view that text is always deeply structured and never identical to a mere sequence of signs or words on a page, this is the definition that I will use.

Transcription: The process of what would traditionally be called transcription depends on the type of the available witnesses. I assume here the situation of the intercepted witnesses, mainly HTML. It is clear that when editing a web site the verb 'transcribe' is no longer applicable, because of the volume and electronic properties. However, the text fragments, the data and their organisation have to be culled from the various witness files. Automated processes will determine which book a review is about, what the text of the review is and which rating has been assigned. In this context, a better word than transcription is conversion: the process converts the information implicit in the scraped witnesses into explicit information. If we want to stress that this is a process of enrichment, we can say the process up-converts information into the explicit textual structure that underlies the edition. We should understand that the fact that conversion is automated does not make it objective or perfect.

Variation is related to the dynamics of the site. It cannot exist in what we have called a synchronic edition of a website. In a diachronic edition variation will be mostly of the structural kind: added or deleted responses, books and users. Changes in the embedded text fragments will occur but will be infrequent. In an extended notion of text, this structural variation can still be called textual, but an apparatus with readings will not be the proper way to record it. Researchers' queries will determine when the edition will show all reviews or only a selection valid at a certain point in time. Variation in a diachronic site model is essentially additive; while in many editions variants represent alternative readings and, depending on editorial orientation, may be seen as mutually exclusive, in the diachronic edition model a response no longer present in a later state of the site is not thereby invalidated.

5 A conceptual model of the edition

An edition of a book discussion site would probably consist of several layers. I suggest the following layers would be useful; not all are necessary:

1) An archive containing witnesses for the entire site at a specific date or possibly at multiple dates. This is necessary because, as we have seen, the sites can easily change or disappear and we need a firm basis for further edition layers.
2) A functional read-only reconstruction of the site. The reconstruction need not replicate the details of the interface, but it should include the content, main navigation paths and a search facility. In a sense, this is the edition's 'reading text'. A clean reconstruction is, however, bound to a certain date. A synchronic edition, which contains information from multiple dates, might offer multiple clean reconstructions. A reconstruction that blends information from multiple dates would probably not be a good idea.

   A reconstruction will contain the front pages, help pages and other site-level information that will probably be absent from a structured database of the site's content (see next layer). A reconstruction of a site is not unproblematic, however. It was noted above that the display of web pages depends to some extent on user hardware and software and potentially other information the site has about the user. What platform do we see as canonical? Is that the desktop view, the tablet view or the smartphone view?

   To some extent, the reconstruction may be an immediate display of the saved (HTML) witnesses. But in cases where javascript is used to dynamically manipulate the content of the web page, the reconstruction will have to go beyond display of saved HTML and might be based on the reconstruction of the site's database (the next layer).

   The reconstruction layer is the primary access point for researchers interested in 'close reading' access to the collection. Layers 1 and 2 together is what is offered in current web archives, such as the Internet Archive[5] or the web archive of the National Library of the Netherlands.[6] Everything that requires modelling of specific sites' content is clearly impossible for a general-purpose web archive. Search facilities too are usually lacking or not functional in web archives.
3) A database of the site's content in terms of the site's structure (or perhaps in terms of an overarching domain model). The database is the result of up-converting the HTML and other witnesses, as mentioned above. It may look like the database that underlies the live sites, and will contain entities such as books, reviews and users.
4) A powerful search and query facility. As well as text-based searches this should allow structured queries that ask for e.g. the number of reviews by month or the books with more than fifty reviews that are rated 4.5 or higher. There might be a simple and an expert user interface. This layer would probably be the primary access point for researchers interested in 'distant reading' access to the material.
5) The edition might include a number of curated paths through the material, for example for those who want to get to know the collection, those who are interested in the reception of an important work, or those who want to be introduced to certain prolific or otherwise interesting reviewers.
6) The edition might offer further derived formats, such all texts organized as a linguistic corpus, queryable using the Corpus Query Language (CQL), or all

[5] https://archive.org
[6] http://webaccess.kb.nl:8080/archived/. Unfortunately only accessible from the premises of the National Library.

structural information in terms of a Linked Open Data graph, queryable using the SPARQL query language.

7) An Application Programming Interface (API) for those that want to research the collection with their own tools; ideally, this would be the interface to the database that the edition itself also uses.
8) An access control layer to restrict access to researchers only and, perhaps, to anonymize information. This is a layer that one would rather do without, but in cases where it is infeasible to ask the rights' owners for permission it may be necessary.
9) An introduction to the edition explaining the rationale for the edition and the decisions made in its creation, including access to the software used for creating the edition.

6 Editing online book discussion?

To make online book discussion accessible to researchers can certainly be called, in Patrick Sahle's convenient formula, 'erschließende Wiedergabe' of historical documents (Sahle 2016 23), and so could count as editing. The preparation of such a collection would then be an exercise in textual scholarship. The question is whether using these terms in this context is clarifying or obscuring the nature of such a project.

It is not idle to invoke the term 'textual scholarship' in the context of born digital documents. Kirschenbaum (2016b) and Ries (2017) use the concept in the context of the analysis of traces of the writing process left on computer hard disks that can shed light on the genesis of literary documents. Hiller (2014) discusses textual scholarship and editing of computer program source code. Since writing a popular blog has become a way of landing a book contract, we are seeing more and more books with a partially online *avant-texte* whose genesis cannot be studied without taking into account online text. It is clear that sooner or later textual scholarship will have to deal with texts that were born online.

I also take it that there is no need to argue extensively that collecting online discussions is useful or even necessary to historians and to researchers more generally. Political historians will need access to blogs and Instagram to understand our election process (Kirschenbaum 2016a) and cultural historians will need access to 'mum blogs' to understand our concept of motherhood (Friedman 2010). 'The web […] is a key resource for understanding human behavior and communication,' write Dougherty and Meyer (2014), and conclude 'it is necessary to begin to take web archiving much more seriously as an important element of any research program involving web resources.'

There are many reasons why online book discussion should be included among these valuable sources. When, in a sixteenth- or seventeenth-century volume, we encounter contemporary underlinings or comments, we are glad because they help us assess how that book was read. Online book discussion can fulfil that function for today's literature. Its availability can help move literary scholarship into a reader-oriented direction. Though it cannot be considered representative of all readers, its volume makes sure that a wide spectrum of opinion is represented. Its volume and its being digitally available also make it possible to do statistical analyses correlating properties of the discussed books, of the reviewers and of the language that the reviews use.

That quantitative research is indeed possible on the basis of online book discussion has been shown by the example of marketing research based on Amazon book reviews (e.g. Chevalier and Mayzlin 2006; Danescu-Niculescu-Mizil et al. 2009). Other topics of research for which online book discussion has served as input includes support for readers advisory services in the public library (Ridenour and Jeong 2016; Spiteri and Pecoskie 2016) and for social book search from an information retrieval perspective (Koolen et al. 2013). For other research around online book discussion, see Boot (2013).

Now, if online book discussion is relevant to scholarly research, and research is actually being done on it, the creation of the research collection is also a scholarly responsibility. It would be naïve to assume that just because we download large quantities of data that is online available, we need not be aware of the choices and limitations embedded in the process. Social scientists have, earlier than humanities scholars, reflected on the dangers of naïve use of social media data. Crawford and boyd (2011) warn that big data claims about objectivity and truth are misleading. They argue that decisions about what will be measured, about data cleaning, about how to handle errors, how to account for unavoidable bias, are all subjective and interpretive and therefore need scholarly reflection. To relegate these decisions to programmers merely means that the scholarly responsibility shifts, not that it disappears (Van Zundert 2015). For the field of history, Hoekstra and Koolen (In preparation) write 'Our main point is that data interaction should be seen as an integral part of doing research. There should be no more room for the sentiment that after the 'data stuff' has been done, the researcher can start doing 'real research'. The data stuff is real research.' What they could have added is: the 'data stuff' is thus a researcher's responsibility, whether it is a programming researcher or a researching programmer.

Perhaps the most important of these scholarly tasks is the modelling activity that is required for the creation of any database, scholarly collection or edition. Modelling was argued to be an essential activity in digital humanities scholarship by McCarty (2005). Pierazzo (2015 11) identifies modelling as 'the key methodological structure of digital editing.' Some of that modelling work for a collection such as the one that I envisage has been described in the previous sections of this article. For a model of the content of such a collection I suggest an entity-relationship model with embedded text fragments. The main entities in the model will be book responses, books and users. For a fuller discussion we have no place here.

Let us use Elena Pierazzo's elaboration of Sahle's definition for a scholarly digital edition here: 'an interpretative representation of historical documents which encompass a combination of primary sources surrogates, edited text(s), and tools to exploit them' (Pierazzo 2015 214) and let us also assume that in this definition the 'edited text' clause does not introduce any new requirements on the process and merely means: the text being presented in this edition. We can now conclude that the collections that I have discussed in this article fulfil all the requirements in this definition and can be considered scholarly digital editions.

Unease with this result may be due to what this definition does not state, but what is true for most editing projects: texts that are being edited are usually canonical works, or works by canonical writers (and the act of editing contributes to that canonical status). The effort that a scholarly edition requires can only be justified for texts that are truly important: the laborious transcription and checking of transcriptions, the careful inspection of variants (if applicable), the various processes of normalisation, perhaps the conjectures aiming to restore corruptions, all of these rest on the assumption that the edited text is a text that

counts. The other side of the coin is of course that, given the limited availability of scholarly time and money, a decision to spend so much time on editing a work is *eo ipso* a decision not to edit so many other works that might also be interesting for researchers.

The digital revolution is changing this situation in a number of different, but related ways. First of all the advent of web 2.0 has increased the availability of digital texts (in computer science terms, 'user generated content'), clearly not all of them masterpieces, but still interesting for research purposes. Second, these texts are online and electronically accessible; their volume necessitates and their being electronically accessible facilitates digital transcription (conversion). This reduces the time and effort needed to include a text in a research collection. Third, the advent of tools for distant reading (such as topic modelling and sentiment analysis, to mention the best known examples) is making it possible to detect patterns in collections of ten thousands of texts that would have been impossible to discover for even the most industrious and perceptive of human readers. This implies that such a collection also becomes much more useful than it would have been without these tools.

Nevertheless, if the (computational) edition process that I propose is so different from what used to be the editor's labour, if the individual texts that would be collected are of less importance than most traditional objects of editing, if our attention to individual words and sentences is necessarily minimal, it is a legitimate question to ask whether it makes sense to call the result of this undertaking an edition. Pierazzo might have added (but did not) a clause that the edition should result from a careful (manual?) examination of all relevant text fragments.

7 Conclusion

In the end, it may not matter very much what we call a collection such as the ones we have been discussing. They are editions in the sense of Pierazzo's definition, although they are so different from most current editions that it feels strange to call them so. They are also like web archives, in that they contain saved web pages and offer a way browsing web sites based on a date in the past, although they offer much more. They are thematic research collections, but that is a rather unspecific term. What does matter is that we begin to think about how we edit or collect texts that share some of the properties of the genre of online book discussion: born-digital online texts from transient, living, deeply-networked and large multi-media sites. For an edition of these, we'll need to create our own witnesses, extend our notion of text, and replace the notion of transcription by upward conversion. An edition, if we may use that name, should be based on a model of the site and provide an archive of witnesses, access to a functional replica of the site, contain a database(−like) representation of the structured content of the site, offer various ways of tool-based access and finally explain the choices that were made in creating the edition.

The task of providing trustworthy text to researchers is relevant in many scholarly fields and can take many forms. Its main concerns, however, are the same everywhere: the respect for the source, the trustworthiness of the text, the transparency of the procedures that were applied, and the flexibility of its output. Textual scholars are bound to encounter ever more online texts and online text collections, and so will scholars from other disciplines. They will have to learn how to handle these texts. In fields such as media studies or internet studies, collections of online material are, of

course, nothing new. Still, they should be able to learn something from the principled and conscientious approaches developed in textual scholarship, just like textual scholars can learn something new for their field from the study of online text.

References

Boot, P. (2013). *The desirability of a corpus of online book responses*. Paper presented at the second workshop on computational linguistics for literature. Retrieved from https://aclweb.org/anthology/papers/W/W13/W13-1405/. Accessed 6 May 2019.

Boot, P. (2017). A database of online book response and the nature of the literary thriller. Paper presented at Digital Humanities 2017. https://dh2017.adho.org/abstracts/208/208.pdf. Accessed 6 May 2019.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research, 43*(3), 345–354.

Crawford, K., & Boyd, D. (2011). Six provocations for big data. Paper presented at A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society. Retrieved from https://ssrn.com/abstract=1926431. Accessed 6 May 2019.

Danescu-Niculescu-Mizil, C., Kossinets, G., Kleinberg, J., & Lee, L. (2009). *How opinions are received by online communities: a case study on amazon. com helpfulness votes*. Proceedings of the 18th international conference on world wide web. New York: ACM pp. 141–150.

Daniels, M. (2016). *"Als ik realisme wil ga ik wel een uur uit het raam staan kijken." Een kwalitatieve en kwantitatieve analyse van Nederlandse lezersrecensies op Bol.com en Goodreads.* Radboud Universiteit, Nijmegen.

Dougherty, M., & Meyer, E. T. (2014). Community, tools, and practices in web archiving: The state-of-the-art in relation to social science and humanities research needs. *Journal of the Association for Information Science and Technology, 65*(11), 2195–2209.

Friedman, M. (2010). On mommyblogging: Notes to a future feminist historian. *Journal of Women's History, 22*(4), 197–208.

Hiller, M. (2014). Diskurs/Signal (II). Prolegomena zu einer Philologie digitaler Quelltexte. *Editio, 28*(1), 193–212.

Hoekstra, R., & Koolen, M. (2018). Data scopes for digital history research. *Historical Methods52*(2), 79-94.

IFLA. (2015). In C. Bekiari, M. Doerr, P. Le Bœuf, & P. Riva (Eds.), *Definition of FRBROO : A conceptual model for bibliographic information in object-oriented formalism*. https://www.ifla.org/publications/node/11240

Jessen, A. (2016). *Lezen als sociale activiteit: van leesgezelschap tot online lezerscommunity.* Radboud University Nijmegen, Nijmegen.

Kirschenbaum, M. G. (2016a). Track changes. A literary history of word processing: Harvard University Press.

Kirschenbaum, M. G. (Producer). (2016b) The Transformissions of the Archive: Literary Remainders in the Late Age of Print. Retrieved from https://www.youtube.com/watch?v=6TuA4dkRegQ. Accessed 6 May 2019.

Koolen, M., Kazai, G., Preminger, M., & Doucet, A. (2013). *Overview of the INEX 2013 social book search track.* Paper presented at the Fourth International Conference of the Cross-Language Evaluation Forum, CLEF 2013.

McCarty, W. (2005). *Humanities computing*. Basingstoke: Palgrave Macmillan.

Palmer, C. L. (2004). Thematic research collections. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A companion to digital humanities* (pp. 348–365). Oxford: Blackwell.

Pierazzo, E. (2015). *Digital scholarly editing: Theories, models and methods*. Farnham & Burlington: Ashgate.

Price, K. M. (2009). Edition, Project, Database, Archive, Thematic Research Collection: What's in a Name? *DHQ: Digital Humanities Quarterly,* 3(3). http://www.digitalhumanities.org/dhq/vol/3/3/000053/000053.html. Accessed 6 May 2019.

Ridenour, L., & Jeong, W. (2016). Leveraging the power of social reading and big data: An analysis of co-read clusters of books on Goodreads. *IConference 2016 Proceedings*.

Ries, T. (2017). The rationale of the born-digital dossier génétique: Digital forensics and the writing process: With examples from the Thomas Kling archive. *Digital Scholarship in the Humanities.* https://doi.org/10.1093/llc/fqx049. Accessed 6 May 2019.

Sahle, P. (2016). What is a scholarly digital edition? In M. J. Driscoll & E. Pierazzo (Eds.), *Digital scholarly editing*. Cambridge: OpenBook Publishers.

Spiteri, L. F., & Pecoskie, J. (2016). Affective taxomonies of the reading experience: Using user-generated reviews for readers' advisory. *Proceedings of the Association for Information Science and Technology, 53*(1), 1–9.

Thelwall, M., & Kousha, K. (2016). Goodreads: A social network site for book readers. *Journal of the Association for Information Science and Technology, 68*(4), 972–983.

Van Zundert, J. J. (2015). Author, editor, engineer—Code & the rewriting of authorship in scholarly editing. *Interdisciplinary Science Reviews, 40*(4), 349–375.

Vanhoutte, E. (1999). Where is the editor? Resistance in the creation of an electronic critical edition. *Human IT, 3*(1) https://humanit.hb.se/article/view/226/299. Accessed 6 May 2019.