The Case for Scholarly Editions

Krista Stinne Greve Rasmussen, Kim Steen Ravn, Jon Tafdrup and Katrine Frøkjær Baunvig [1]

[1] *Center for Grundtvig Studies, Aarhus University, 8000 Aarhus C, Denmark*

Abstract

This article wish to make a case for scholarly digital editions (SDE's). SDE's can help to put the computational humanities into the center of humanistic scholarship. Large scaled projects pave the way for easy access to historical documents, but small, curated digital editions, strenuously enriched by philologists, will be the key player in the process of integrating computational humanities into traditional scholarship; first and foremost because this material is clean, reliable, and flexible, secondly, thorough markup leaves it open to comprehensive, fine-grained, hermeneutically complex explorations.

Keywords

N.F.S. Grundtvig, Scholarly Digital Edition (SDE), users, readers, OCR, Humanities computing, data, knowledge sites, data validation, data reliability, data infrastructure

1. Introduction

Big and dirty corpora stand in opposition to smaller, cleansed and philologically digested data-sets that ensure research validity. For a computational cultural historian the former can, because of their large scale, be used to paint a blurry background of semantic trends in a given period, despite the 'dirtiness' of the data. However, the often smaller, but purer and philologically handled data helps to paint the foreground, which is more accurate and greater in details.

This article will argue that 'big and dirty' cannot stand alone in fully consolidating computational humanities. As said, the somewhat fuzzy data in large scale digitization is like a the background of a painting, where details fade, while data provided by SDE's make it possible to demonstrate details in the foreground. Furthermore, the SDE itself provides tools for more usable investigation and reading. Therefore, it is time for making a case for scholarly editions.

2. Big and dirty

The digitization of texts often transforms large collections of texts into digital corpora by means of OCR (Optical Character Recognition). It is a rather straightforward procedure and it is the foundation for valuable mega-scaled projects such as the Google Books and for meso-scaled projects such as *Mediastream (Mediestream),* the digitization of Danish newspapers published from 1666 and onwards from the Danish Royal Library in Copenhagen. The materials have received only cursory proofreading (or perhaps none at all), and it is well-known that they contain variating degrees of 'dirt' in the form of transcription errors.

However, the usefulness of large corpora is evident despite the risk of transcription errors. At Center for Grundtvig Studies Katrine Frøkjær Baunvig (KFB) has carried out research that both use big data, as they are [1] and research that combines the use of dirty data with manually cleansing. This way it is

The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15-18, 2022

EMAIL: ksgr@cas.au.dk (K.S.G. Rasmussen); teoksr@cas.au.dk (K.S. Ravn); teotaf@cas.au.dk (J. Tafdrup); baunvig@cas.au.dk (K.F. Baunvig)
ORCID: 0000-0003-3424-2334 (K.S.G. Rasmussen); 0000-0001-7600-7645 (K.S. Ravn); 0000-0003-3696-3899 (J. Tafdrup); 0000-0003-3424-2334 (K.F. Baunvig)

possible to do both distant and close reading within the same study, like the study, that uncovered the rise of the demi-goddess ‘Dana’ in Danish nineteenth-century newspapers 1800–1870 [2].

In this study KFB accepted that the material contained OCR dirt, which was compensated with hours of manual cleaning, annotation and coding by student assistant Stine Kylsø Pedersen. The study was based on data made available through the Danish Royal Library’s digital archive *Mediestream* [3]. The Danish Royal Library have scanned 35.464.209 Danish newspaper and periodical pages published between 1666 and 2013; by way of OCR software this material is made searchable in an interface allowing for date delineations. The result was a simple list-generation conveying the number of occurrences of ‘Dana’ in Danish newspapers during the period in question. Such a list allows for a distant reading clearly demonstrating a steady increase in mentioning of ‘Dana’, however, it was the manually coding and control of false positives that gave the most promising results. This was possible because facsimiles are available alongside the transcripts. Every true occurrence of ‘Dana’ was tagged to gain an overview of semantic context and genre. A time-consuming, but rewarding endeavor that made it possible to outline the coming of age of a national spirit.

Other studies use big data without this kind of time consuming manually labor, but the conclusions regarding the data are the same:

> Such corpus material combined with the ready-made display tools they offer are seductively easy to use. With no access to the back-end – to the underlying text-material of the datasets – in effect, they are black boxes suited only for heuristic purposes and must be handled with a great deal of hermeneutic caution. But when one is disciplined in terms of limiting the stakes of the search (when one does not rely on them for answers to central research questions), these big, dirty and obscure datasets can be the foundation of quite indicative sketches [4].

Big data is useful to point out trends: like the increasing popularity of i.e. ‘Dana’ – or as KFB has done in another study, to demonstrate the relative importance of Grundtvig and Danish philosopher Søren Kierkegaard [5]. Using first *Mediestream* as a tool the investigation seek out the occurrences of their (in Danish context infrequent) surnames in the national newspapers. That includes the timespan from their deaths respectively until today.

Historically it is no surprise that Grundtvig dominated the Danish public discourse throughout the entire period. Grundtvig was a public character and a politically engaged writer, Kierkegaard the opposite. As one would expect peaks in occurrences are found in years of jubilees of birth and death. Furthermore, a Grundtvig revival appears during the Second World War, while Kierkegaard’s relevance seems stable from 1920s onwards.

The examination is also given an international perspective when the corpus of Google Books is included. In 2019 when the investigation was conducted 40 million books had been scanned. The frequencies reveals that Kierkegaard experiences a breakthrough in the first third of the 20th century in the English, German and French corpora. From the middle of the century and ongoing, he suffers a significant decrease in German and French while his relevance has increased remarkably in the English corpus since the late 1980s. Grundtvig is virtually absent in these corpora. He does not seem to have established himself as a significant frame of reference in theses languages.

The study demonstrates how large datasets can be used to substantiate a general notion. In a Danish context, the findings confirm what would have been predicted. However using the easy accessible ready-made display tools offered by Google Books the dataset can be taken as a point of departure to a more substantial analysis of a given matter.

3. Small and pure – *Grundtvig’s Works*

N.F.S. Grundtvig (1783–1872) was a poet, pastor, politician, and romanticist. He is regarded as one of the most (if not *the*) central figure in the nineteenth-century Danish nation building process, as well as in the construction of a modern Danish Christianity [6]. A consortium including members of the Danish parliament decided in the late 2000s to pave the way for the creation of a scholarly digital edition of Grundtvig’s approximately 1.073 published works. The edition is available to Danish citizens free of charge [7].

A scholarly edition of *all* of Grundtvig's published works was wanted and needed within scholarship, and it had never been done before [8]. Center for Grundtvig Studies is also committed to research and dissemination. Therefore, we have made all our material available to researchers.

Grundtvig was an active, publishing writer during a period of 68 years (1804–1872). His printed works amount to approximately 37.000 standard pages of 2400 units per page. The data set has a median document size of four pages and contains 3.968.841 word tokens distributed over 115.240 word-types. As mentioned above the corpus consists of roughly 1000 works, the number declines a little as we make our through the oeuvre as some works are left out and others added. Our definition of a work is any text that Grundtvig made public available through printing.

By way of OCR technology, first editions of each text have been 'translated' from a TIF-format (Tagged Image File Format) into an XML (Extensible Markup Language) document, manually annotated following the TEI (Text Encoding Initiative P5) standards. As anyone familiar with digitization knows, the labor-intensive cleansing and annotation of the raw and oftentimes somewhat dirty OCR results is crucial. We collate the texts manually three times as part of our textual criticism. This work is carried out by student assistants and Grundtvig-specialized scholarly editors – philologists trained in fields relevant for the domestication of Grundtvig's prose, such as (obviously) nineteenth-century Danish but also Old Norse, Greek, Bible Studies, hymnology, romantic philosophy, eighteenth-century historiography etc. Furthermore, each work is provided with contextualizing introductions and glossaries. Finally yet importantly, the scholarly edition contains various contextualizing materials, including indices on persons, places, titles and mythological and Biblical references. These indices are of great value and used as local commentaries within the text.

The edition we are providing is a genuine scholarly digital edition (SDE). A scholarly edition is to put it short: "the critical representation of historic documents" [9]. The critical representation is ensured by methodologically consistency and textual criticism. This is the same for both printed and digital editions, but the latter are furthermore defined as follows: "Scholarly digital editions are scholarly editions that are guided by a digital paradigm in their theory, method and practice" [10]. An edition like *Grundtvig's Works* that has been conceived for and designed to be fully digital is of course guided by a digital paradigm, even though many questions of textual criticism are the same, i.e. what text to use for the edition (we use first editions, not the last printed). However, any scholarly edition can be guided by a digital paradigm in numerous ways, discussions of modelling editions and what the digital in 'digital scholarly editions' mean are widespread.

Peter Shillingsburg has promoted the idea of the 'knowledge sites'. Others talk about the "work-site" [11] and "fluid, collaborative and distributed editions" [12]. Common to these definitions is a wish for digital scholarly editions to be broader in scope. Shillingsburg defines the knowledge site as a site where:

> textual archives serve as a base for scholarly editions which serve in tandem with every other sort of literary scholarship to create knowledge sites of current and developing scholarship that can also serve as pedagogical tools in an environment where each user can choose an entry way, select a congenial set of enabling contextual materials, and emerge with a personalized interactive form of the work […], always able to plug in back for more information or different perspectives [13].

The idea of the knowledge sites and relating notions of scholarly editions builds on a concept of the digital edition as a radical break with traditional editions and notions of the limitations of the printed book. The digital scholarly edition can definitely be modelled in numerous ways, and this is not the place to discuss edition designs. Rather, the point here is that scholarly digital editions are part of important "knowledge sites", not understood as websites or kinds of editions, but as, properly speaking "knowledge environments" with scopes that cannot be anticipated by the philologists creating the edition, nor included within the edition. The knowledge environment we provide is not at knowledge site in the meaning proposed by Shillingsburg. It is not interactive, but it has features the reader can use to better read and investigate the works.

A crucial difference between print and digital is the shift in the reader's possibilities to read, use and engage with the edition. Krista Stinne Greve Rasmussen has defined this as three different reader roles made possible by the digital paradigm: reader, user, co-worker [11]. These roles designate difference types of engagement with an edition. The possibility of using the data for research within humanities computing is an obvious way of engaging with an edition as user. However, the studies mentioned in this article does not properly speaking engage with the edition per se, they engage with data provided by the edition. This is an obvious but crucial benefit of a digital scholarly edition; the data it provides is pure and clean:

> small, curated digital editions, strenuously enriched by philologists, will be the key player in the process of integrating CH [Computational Humanities] into traditional scholarship; first and foremost because this material is clean, reliable, and flexible: Thorough markup leaves it open to comprehensive, fine-grained, hermeneutically complex explorations [12].

Word embedding is one method that relies on pure data as demonstrated in the article, "Mermaids are Birds. Embedding N.F.S. Grundtvig's Bestiary" in this conference volume [16]. For the study they, furthermore, took advantage of one of the indices (the "Mythological register" in *Grundtvig's Works*), but again, not directly as a user of the edition, but as users of the edition as a 'knowledge environment' so to speak.

4. The case for scholarly editions

The case we want to make for scholarly editions here is perhaps simple, but important: Scholarly digital editions provide reliable texts of important works, easily accessible online (and in our case free of charge) and conveyed to the public with contextualizing materials. This is to some extent the raison d'être of a scholarly edition and the reason for funding a project like ours. For the digital humanities however, making a scholarly digital edition also means providing a clean and pure data set, and the purer the more one can compute.

The intended audience and funding is to some degree codependent. An expensive edition funded with public support such as *Grundtvig's Works* must in a Danish context *per se* be intended for a broad audience supporting both readers and users. Creating and participating in knowledge environments for dedicated scholars within computing humanities is not only beneficial for the scholar (be it Grundtvig-scholars, literature- and culture-scholars, historians, theologians, anthropologists etc.), but also for the student at all levels. A clean, critically created corpus with relevant meta-texts (hyperlinks) is high-quality digitization, than can be used both quantitatively and qualitatively as opposed to large, dirty corpora that are mainly for quantitatively use, unless they are manually processed by the user. The manpower put into scholarly digital editions means clean, reliable, and flexible material with thorough markup that leaves it open to comprehensive, fine-grained, hermeneutically complex explorations.

5. References

[1] Baunvig, K. F.: "A Computational Future? Distant Reading in the Historical Study of Religion", In: M. Freudenberg, T. Karis, M. Rademacher, Jens Schlamelcher, F. Elwert (Eds.), Twelve Years of Studying Religious Contacts at the KHK: Stepping Back and Looking Ahead. Dynamics in the History of Religions, Leiden: Brill (forthcoming).

[2] Baunvig, K. F.: "Fictional Realities of Modernity: The Fantastic Life of the Demi-Goddess Dana in the Emerging Nation State of Denmark", in: L.K. Martinsen, S. Bønding, P.-B. Stahl (Eds.) *Mythology and Nation Building in the Nineteenth Century: N.F.S. Grundtvig and His European Contemporaries*, Aarhus University Press, Aarhus (2021), pp. 97–134.

[3] Det Kgl. Bibliotek: Mediestream. URL: http://www2.statsbiblioteket.dk/mediestream/avis

[4] Baunvig, K. F.: "A Computational Future? Distant Reading in the Historical Study of Religion"

[5] Baunvig, K. F.: "A Computational Future? Distant Reading in the Historical Study of Religion"

[6] Holm, A.: The Essential N.F.S. Grundtvig. Copenhagen, Filo (2019).

[7] Center for Grundtvigforskning: Grundtvigs Værker. København/Aarhus (2010-). URL: http://www.grundtvigsværker.dk/

[8] Kondrup, J.: ”Udgivelse af dansk litteratur 1900-2018”, in: J. Kondrup (Ed.), Dansk Editionshistorie, vol. 3, Udgivelse af dansk litteratur, Museum Tusculanums Forlag (2021), pp. 295-476; p. 440

[9] Sahle, P.: “What is a Scholarly Digital Edition?”, in: M.J. Driscoll, E. Pierazzo (Eds.), *Digital Scholarly Editing: Theories and Practices,* Open Book Publishers, Cambridge (2016), pp. 19-39; p. 23. URL: http://books.openedition.org/obp/3397.

[10] Sahle, P.: “What is a Scholarly Digital Edition?”, p. 28

[11] Eggert, P.: ”Text-encoding, Theories of the Text, and the ‘Work-Site’”, Literary & Linguistic Computing 20, 4 (2005), p. 433 f.. DOI:10.1093/llc/fqi050.

[12] Robinson, P.: ”Where We Are with Electronic Scholarly Editions, and Where We Want to Be”, Jahrbuch für Computerphilologie 4 (2004).

[13] Peter Shillingsburg, P.: From Gutenberg to Google: Electronic Representations of Literary Texts, Cambridge, Cambridge University Press (2006), p. 88.

[14] Rasmussen, K.S.G.: “Reading or Using a Digital Edition? Reader Roles in Scholarly Editions”, in: M.J. Driscoll, E. Pierazzo (Eds.), *Digital Scholarly Editing: Theories and Practices,* Open Book Publishers, Cambridge (2016), pp. 119-133. URL: http://books.openedition.org/obp/3397.

[15] Baunvig, K. F.: “A Computational Future? Distant Reading in the Historical Study of Religion”

[16] Katrine F. Baunvig & K.L. Nielbo: Mermaids are Birds. Text Mining N.F.S. Grundtvig's Bestiary. In *Digital Humanities in the Nordic and Baltic Countries Conference* (DHNB 2022)