Storage over Rendition. Call for a Sustainable Infrastructure in the Digital Textual Heritage Sector with a Particular Interest in Digital Scholarly Editions

Katrine F. Baunvig\textsuperscript{1,2,3}, Krista S. G. Rasmussen\textsuperscript{1,3}, Kirsten Vad\textsuperscript{1,3} and Per Møldrup-Dalum\textsuperscript{2,3}

\textsuperscript{1}Center for Grundtvig Studies (Aarhus University), Jens Chr. Skous Vej 3, 8000 Aarhus C, Denmark
\textsuperscript{2}Center for Humanities Computing, Jens Chr. Skous Vej 4, Building 1483, 4th floor, 8000 Aarhus C, Denmark
\textsuperscript{3}Aarhus University, Nordre Ringgade 1, 8000 Aarhus C, Denmark

Abstract
A significant amount of human and pecuniary resources has gone into the production of the long line of digital scholarly editions that within recent decades have sprung to life in Scandinavia, in the Baltic region, as well as in the rest of Europe. Notwithstanding the heritage perspective, the sector is paradoxically characterized by a presentist preoccupation with instant results – first and foremost with the rendition of the given data set. Concerns for long-term data management perspectives – that is: interest in the post-production afterlife of the data – is relatively meagre. This goes for project managers, for project host institutions, and for the research foundations financially supporting the projects. So, despite harmonizing initiatives at production level and pre-edition compilation initiatives, such incentives promote a situation of insulated digital scholarly editions focusing on unique URLs and distinctive qualities of the given material. This hinders project synergy in the production phase. Moreover, it hinders the construction of long-term and sustainable data management solutions. To remedy this situation, we propose a clear division of labour between the tasks of data production, of data rendering, and of data storing. This division should ideally be sought at an institutional level. This will secure the accumulation of know-how in teams refining the respective workflows. In addition, we encourage private and public foundations to bolster this infrastructure by making project compliance a criterion for funding.

Keywords
Scholarly Editions, Digital Scholarly Editions, Cultural Heritage Sector, Data Storage, Data Preservation, Data Production, Data Rendition, Research Infrastructure
1. Introduction

This paper presents little short of an infrastructural vision for the digital cultural heritage sector with particular interest in the area of digital scholarly editions [DSE]. We hope to convince stakeholders in the Nordic region that it is high time for a division of labor within the field. That is, we hope to persuade fellow project managers, project host institutions, GLAM personnel, politicians, and the funding agencies supporting the projects, to join forces and aim for two goals.

1) First, we have to break the main DSE tasks into three (relatively) individual procedures: a. The first concerning to the pre-production, production, and post-production of a given material. b. The next regarding the rendition of this material to given current target groups. c. The third pertaining to the secure and long-term storage of the data material.

2) Meeting UNESCO’s call for open science and a sustainable research data infrastructure, we aim to coordinate, streamline, and bolster the data storage task at a regional level – in a FAIR manner. The FAIR principles for scientific data management and stewardship seek to secure the Findability, Accessibility, Interoperability, and Reuse of digital assets.

The need for moving the field in this direction is both internally and externally conditioned. Internally: the DSEs have within the last couple of decades accumulated a significant amount of general know-how and routine that we, due to the somewhat fragmented nature of the project landscape dominated by a silo-mentality, still have not been able to reap the potential synergy rewards from. Externally: the push for an open science infrastructure and for FAIR data management plans has intensified within the last five years. In other words, the time for action is thus now: On the one hand the DSEs have reached maturity as a scholarly field and are now able to obligate cross-project coordination, on the other hand, data management procedures have improved radically while the data management impetus and ethos have intensified correspondingly.

In sum: Notwithstanding the heritage perspective, the sector is in other words paradoxically characterized by a presentist preoccupation with instant results – first and foremost with the rendition of the given data set. There seem to be somewhat standardized operations in place for the production and rendition of digital scholarly editions, however, solutions for long-term data management perspectives – that is: the postproduction afterlife of the data – are as of yet unconsolidated. This is the situation among project managers, among project host institutions, and among the research foundations financially supporting the projects. So, despite harmonizing initiatives at production level and pre-edition compilation initiatives (e.g., www.litteraturbanken.se), the incentive structure still promotes a situation of insulated digital scholarly editions focusing on unique URLs and distinctive qualities of the given material. This hinders project synergy in the production phase. Moreover, it hinders the construction of long-term and sustainable data management solutions.

2. The History of Scholarly Editions

Scholarly editions of text corpora of various natures have a long and prolific history. Critical Bible studies and studies of the works of the philosophers of Roman and Greek Antiquity –
established as a scholarly practice in the Renaissance, refined in the Early Modern Reformation period, and accelerated throughout the Enlightenment Era – are the deep roots of the endeavors. Through the course of the 19th century, the procedures of preserving and rendering given sets of text as close to the originals as possible were, however, broadened to include phenomena in vogue: the current ‘Genius’ and the ‘Hero’. In other words: the authorships or œuvres of great men whose lives and deeds were still fresh in collective memory. This was the period of the emerging nation-states’ construction of cultural identities and formation of a cultural heritage canon; scholarly editions aided in this process and as a result scholarly editing gradually grew into a field proper (p. 231-266) [1]. Through the 20th Century, the field consolidated into two main areas: classical philology, concerned with codices, and new philology centered on printed books. [1]. Within the Nordic countries, these two fields operate somewhat independently.

3. The History of Digital Scholarly Editions

Scholarly editing has always been pivotal for the development of the so-called Digital Humanities, whose origin is often dated to the year 1949 when Roberto Busa, in collaboration with IBM, began working on his concordance of the works of Thomas Aquinas: Index Thomasticus. Busa’s work is widely recognized as the birth of the field ‘humanities computing’, although his index is not a scholarly edition per se. Index Thomasticus was first published in the 1970s and is now available online. [2]. The first digitization of a canonical Nordic author also formed the basis of an index when the Canadian philosopher Alastair McKinnon, affiliated to McGill University in Montreal, had all of Søren Kierkegaard’s writings transferred on to punched cards in the 1960s. The indices were published in print, and do not now exist digitally. [3] This digitalization later morphed into a digital edition, which was among one of the first in the world. This edition again morphed in to the latest scholarly edition of Kierkegaard’s œuvre: Søren Kierkegaards Skrifter (The Writings of Søren Kierkegaard) [4].

These editions mirror the general development of digital scholarly editions: There is a distinct pre-internet era, in which editions were digitally prepared, but mainly distributed in print (or in some cases CD-ROM). Then follows the internet era in which editions move online, and much theorization and effort was invested in creative rethinking of scholarly editions. The field has consolidated, and the present-day situation is more or less post-internet and occupied with the redefinition of digital scholarly editions that are both sustainable and viable on the long haul of the post-production stage.

---

[1] New philology in this sense only designates the material, other uses of the term ‘new philology’ are material printed (within medieval studies), or focused on native-language sources within indigenous societies. See: (p. 999-1006) [2], [3]


4. Digital Scholarly Editions: Challenges to the Sustainability of the Field

For a long period of consolidation, the DSE have, thus, as a professional community built up and refined DSE routines. It is an obvious fact but not a trivial observation that a significant amount of cultural heritage data has come to life in digital formats within the last decades. It is also an obvious fact but not a trivial observation that a significant amount of human and pecuniary resources has gone into the production of the long line of DSEs that within recent decades have sprung to life in Scandinavia, in the Baltics, as well as in the rest of Europe. The projects piling up on the Berlin-based Institut für Dokumentologie und Editorik’s “A catalogue of Digital Scholarly Editions” (https://v3.digitale-edition.de/vlet-collected-works.html) testify to an overall European trend pertaining to resource allocation. The digitization and computational exploration of cultural textual heritage material attract funding (Rasmussen et al. 2022). This we take as a sign of maturation within the field. Two trends, however, clearly pose a hindrance or a challenge to the sustainability of DSE as a mature field. One is the centripetal power of the canon, the other is the dominating silo-mentality.

4.1. The Centripetal Power of the Canon: The Case of the DSE Grundtvig’s Works

In Danish public discourse poet, pastor, politician, and romanticist N.F.S. Grundtvig (1783–1872) 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: In short, he is regarded as a cultural saint. In scholarly literature, it is widely acknowledged that Grundtvig sought to stimulate the process of assembling a collective Danish emotional consciousness based on 1) a horizontal-contemporary axis incorporating the different strata within the socially heterogeneous “Folk” and on 2) a vertical-historical axis connecting present-day Danes with forefathers and legendary characters. In social historian Benedict Anderson’s words, the emotional fabric intended by this attempted interlacing was an imagined community. Nowadays, Grundtvig’s cultural imprints are acknowledged by most Danes: “N.F.S. Grundtvig founded Danish democracy”; “N.F.S. Grundtvig established the Church of Denmark (folkekirken)”; “N.F.S. Grundtvig is the founder of the Danish school system”; “N.F.S. Grundtvig revived the pre-Christian Nordic tradition”; “N.F.S. Grundtvig is the most important writer of Christian hymns in Denmark”. These are surprisingly recurrent statements in Danish public media, deeming his intellectual activity more culturally important than the work of his world-famous contemporaries Søren Kierkegaard (1813–1855) and Hans Christian Andersen (1805–1875).

Considering Grundtvig’s cultural status, it is perhaps no surprise that a consortium including members of the Danish parliament some 15 years ago decided to pave the way for the creation

---

5This ‘sainthood’ is not a banal, cosmetic analogy. Reverence and quasi-ritual structures have been built around Grundtvig as a ‘Great Dead’ [4]. Grundtvig has a cathedral named after him: the Copenhagen Grundtvigs Kirke; every year his birthday (almost coinciding with the day of his death) is celebrated in Grundtvig-relevant institutions; one such celebration entails the opening of his crypt at the small cemetery Clara’s Kirkegård on the outskirts of the Sealandic town of Køge. Moreover, Grundtvig’s Death (Grundtvigs død) is a commodity – at least it is a recent title in a popular book series by Aarhus University Press written by Grundtvig scholar Jes Fabricius Møller (2019).
of a scholarly edited, digital version of Grundtvig’s published writings, making them available to Danish citizens free of charge. The writings, published within a period of 68 years (1804–1872), amount to 4M word tokens distributed over approximately 37K standard pages.

First editions of each work are OCR prepared and are now being manually cleansed. A crucial step in producing an accurate digital corpus is this labor-intensive cleansing and annotation of the raw and oftentimes somewhat dirty OCR results. This corpus forms the basis of the digital scholarly edition produced by a group of ten Grundtvig-specialized scholarly editors – philologists trained in fields relevant to 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, political history, etc. As part of the edition, this équipe furnishes the individual texts with contextualizing introductions and glossaries. Their work, piling up on www.grundtvigsvaerker.dk, was initially estimated to consume 200 man-years. This prognosis seems to hold: 11 years and 100M DKK (approx. 15M Euro) later, the ten scholarly editors are halfway through the project. Such details are highly relevant when drawing up the contours of DSE – and to some extent also of the Humanities Computing – because the thorough markup is what leaves the data open to the general public as well as to comprehensive, fine-grained, hermeneutically complex scholarly explorations.

Clean, reliable, and flexible but the point here is that it is also highly exclusive. High quality data is, as we have just tried to spell out, burdensome to create in terms of time and funding. AND not every type of material, not every type of authorship will move politicians and research foundations to cover the expenses6 - for the last decades mainly canonical figures and archives have done so. A long line of political and ethical problems nest here.

But gradually funding of corpora of a different nature seems to circumvent this drift towards the canon – or the canon seems to evolve in a more inclusive, socially sustainable direction. If this drift catches on, it would be a great gain for the cultural heritage sector, for humanist scholars, and for the general public.

4.2. The Silo-Mentality dominating the DSE Sector

The early and consolidating days of the DSE sector have been characterized by a long line of individual projects with no or not that much infrastructural coordination. Obviously, personal, and scholarly networks have sought to mitigate the fact that it was every DSE project for itself / that every project ultimately was responsible for each stage in the DSE production line – this fostered what could perhaps be thought of as a silo mentality. Highly different in nature and scope, NNE (Nordisk Netværk for Editionsfilologer) and TEI are two obvious examples of initiatives aiming to enhance coordination among individual projects. Though much great work has been done in, around, and with NNE and TEI, we see them as symptoms of a core problem that there is a lack of harmonization, coordination, or division of labor within the sector. One obvious reason being that each edition is considered a full-scale research project in itself, mimicking scholarly editions of the print era. However, a print and a digital edition are not the same, storage and rendition being the main game changers.

6The Great Unread vs. The Canon has long been a topic in Literary History (Margaret Cohen, The Sentimental Education of the Novel (Princeton: Princeton University Press, 1999), Franco Moretti, “The Slaughterhouse of Literature” Distant Reading (London: Verso, 2013)).
Many steps have been taken to move scholarly editions into the digital paradigm, which has affected the production in almost every step. Editors have learned TEI and XML, and many have even attained the expertise to produce their editions online. Now they struggle with the worries of storage, and many probably spend troublesome hours faced with a field of expertise, they—rightly—feel uneasy to enter, that is long-term preservation of digital, cultural heritage. The difference between print and digital editions is nowhere as prominent as when it comes to rendition and storage. Editors of print editions leave these concerns mostly in the hands of printing houses and libraries. This is still not the case with digital editions, and more often than not rendition and storage end up in the silo of the scholarly edition.

Fully adhering to the digital paradigm means that rendition and storage have to be non-silo endeavours. Research infrastructure and joint publishing tools have been promoted, but hitherto the responsibility for and expertise into long-term rendition and storage have resided with the editor. An approach that is neither sustainable nor viable.

5. Digital Scholarly Editions: A Mature Field

Challenges aside, a cross-sectoral routine or task flow seems to have somewhat stabilized during the last decade. Every project thus experience having to deal with a given material in need of processing by way of, e.g., OCR or Transcribus. The next step running across the DSE field is the manual cleansing and mark-up of the given material carried-out by digital scholarly editors. The result hereof is what is rendered on given URLs for lay and scholarly users to consume; scholarly consumption of the DSE data can, however, also circumvent the rendition and focus alone on the ‘raw’ digitized material. Finally, we have the stage of storage. Here, however, no specific long-term model seems to have stabilized (see Fig. 1).

Researchers have, furthermore, consolidated the field of editorial theory, both at a Nordic and European level, and have shared forces in the constructions of highly elaborate digital scholarly editions on the internet. In this sense it is a mature field that is steering towards a new stage in the history of digital scholarly editions: the digital afterlife. Preservation of the current digital scholarly editions that resides on the internet is facing an obvious problem: they are not maintainable in their current design [5]. The steps we need to take in the immediate future are best taken together within the field that is both a mature research field and a consolidated community.

6. Conclusion: Tripartite Division of Labour

We hereby invite stakeholders involved in digital scholarly editing to remedy this situation. We propose to seek binding regional infrastructures articulating and dividing responsibility for a) the production of the data, b) the short-term rendition of the given data sets, and c) for the long-term storage of data in a FAIR manner. The underlying logic is that data storing represents a humdrum operational task with few rewards in terms of potential institutional exposure and public acknowledgement. This explains the slapdash and unambitious solutions available.

---

*Such as i.e. TEI Publisher and CLARIN.*
Figure 1: The DSE Work Loop. Every project will experience having to deal with a given material in need of processing by way of, e.g., OCR or Transcribus. The next step is the manual cleansing and mark-up of the given material carried-out by digital scholarly editors. The result hereof is what is rendered on given URLs for lay and scholarly users to consume; scholarly consumption of the DSE data can, however, also circumvent the rendition and focus alone on the ‘raw’ digitized material. Finally, we have the stage of storage. Here no specific long-term model seems to have stabilized.

Nevertheless, proper storing is the only sustainable argument for the resources going into the production of the digital editions.

In conclusion: We propose a clear division of labour between the tasks of data production, of data rendition, and of data storage. This division should ideally be sought at an institutional and a regional level. This will secure the accumulation of know-how in teams refining the respective workflows. In addition, we encourage private and public foundations to undergird this infrastructure by making project compliance a criterion for funding.

References

