CRITICAL DIGITAL TRANSMISSION THROUGH THE TRANSCRIPTION OF J. A. NENSÉN’S RECORDS
A REVIEW OF THE TOOLS

JACQUELINE VIGILANTI

HÖGSKOLAN I BORÅS
VETENSKAP FÖR PROFESSION

CC  Jacqueline Vigilanti
This work is licensed under the Creative Commons Attribution 4.0 International License.
This thesis aims to understand transcription as a critical transmission practice in the digital medium in order to add to the research surrounding digital library initiatives that make these research materials available and accessible. The project to transcribe the records of Lapland priest J. A. Nensén at Umeå University aims to publish the transcriptions in the cultural heritage database Alvin by Uppsala University Library. Looking at these two Swedish projects can better illustrate the issues surrounding digital tool application and integration. This thesis first explores the transcription practices of a group of scholars transcribing J. A. Nensén’s records. By understanding their model, a better picture of transcription as a scholarly practice can be achieved that can be useful when selecting a digital transcription tool. The study then identified appropriate transcription tools and reviewed the ones most applicable for the project based on the group’s paradigm. It was discovered that the main disparity in transcription practice is the balance between producing a diplomatic transcription and one that is readable. The primary consideration in transcription tool selection was found to be the goals of the project; in this case, those of the transcribers. This determined the tools to be reviewed. These were: FromThePage, Scripto, T-Pen, and Wikisource. All the tools are suitable to the editors, but they still present some obstacles in regards to diplomatic transcription.

Key words: Digital library, digital scholarly editing, critical transmission, cultural heritage material, tool review, transcription

Digitala bibliotek, kritisk digitalisering, kulturarv, transkribering, transkriberingsverktyg
Acknowledgements

I wish especially to thank the group working on J. A. Nensén’s records at Umeå University for the opportunity to explore manuscript transcription more deeply. Most particularly, I wish to thank the transcribers, Stina Karlström and Robert Eckeryd, for their enthusiasm and interest in my efforts before, during, and after the group interview. Kind appreciation is also expressed to Lars-Erik Edlund for his correspondence regarding information pertaining to the Nensén materials.

Much gratitude is extended to Per Cullhed and Krister Östlund at Uppsala University Library for beginning the discussion on Alvin which led to this project. As well, I am thankful for their time and personal interest during my visits and in communication afterwards.

I would also like to thank my supervisor Mikael Gunnarsson for his suggestions and comments, specifically, for his time and patience while I developed the project.

And in a personal and professional capacity, I wish to express true gratefulness to Cecilia Falk, Luleå University of Technology Library, for corrections to the thesis, inspiration as a librarian, and for providing me with the chance to gain the practical experience working at a library in the first place.

Jacqueline Vigilanti
Luleå, September 2015
# Table of Contents

1  Introduction .................................................................................................................. - 2 -  

1.1  “The Inquisitive Vicar” - digitization of J. A. Nensén’s records ................................ - 3 -  

1.1.1  The collection ......................................................................................................... - 3 -  

1.1.2  Aim of transcribing the Nensén records ............................................................... - 4 -  

1.1.3  Practical considerations ......................................................................................... - 4 -  

1.2  Aims of the research ................................................................................................. - 5 -  

1.3  Objectives of the thesis ............................................................................................ - 5 -  

1.4  Research Questions ................................................................................................... - 6 -  

1.5  Limitations .................................................................................................................. - 6 -  

2  Background and Problem Description ....................................................................... - 7 -  

2.1  Uppsala University Library’s digital archive and database ...................................... - 7 -  

2.2  National and international perspective ...................................................................... - 8 -  

2.3  Problem Description ................................................................................................. - 10 -  

3  Conceptual Background and Critical Context .............................................................. - 11 -  

3.1  Definition of transcription ........................................................................................ - 11 -  

3.2  An editorial task in the digital medium .................................................................... - 12 -  

3.3  Theoretical assumptions ......................................................................................... - 12 -  

3.4  Editorial practice facilitated by digital tools ............................................................. - 14 -  

4  Literature Review ........................................................................................................ - 15 -  

4.1  Critical transmission ............................................................................................... - 15 -  

4.2  Transcription ............................................................................................................ - 15 -  

4.3  Transcription tools in research and practice ............................................................ - 16 -  

4.4  Research methodologies ......................................................................................... - 17 -  

5  Methodology ................................................................................................................ - 19 -  

5.1  Research design ...................................................................................................... - 19 -  

5.2  Data collection methods ....................................................................................... - 19 -  

5.3  Participants .............................................................................................................. - 20 -  

5.4  Design of the group interview ................................................................................ - 20 -  

5.5  Identifying transcription tools ................................................................................ - 21 -  

5.6  Methods of data analysis ....................................................................................... - 22 -  

6  Results .......................................................................................................................... - 24 -  

6.1  Group interview ...................................................................................................... - 24 -  

6.1.1  Goals .................................................................................................................... - 24 -  

6.1.2  The material ......................................................................................................... - 27 -  

6.1.3  Practical issues .................................................................................................... - 29 -  

6.1.4  Transcription model ........................................................................................... - 29 -  

6.1.5  Key themes ......................................................................................................... - 34 -  

6.2  Identification of tools ............................................................................................. - 36 -  

6.3  Selection of tools to be reviewed .......................................................................... - 38 -  

6.4  Tool Review ............................................................................................................ - 40 -  

6.4.1  FromThePage ...................................................................................................... - 40 -  

6.4.2  Scripto .................................................................................................................. - 41 -  

6.4.3  Wikisource .......................................................................................................... - 41 -  

6.4.4  T-Pen .................................................................................................................. - 42 -  

6.5  “Keep it simple” ...................................................................................................... - 43 -  

6.6  “That’s the principle of it all: Diplomatic” ............................................................. - 47 -  

6.7  Markup .................................................................................................................... - 49 -  

6.8  Tool transcription models ...................................................................................... - 52 -  

7  Discussion .................................................................................................................... - 54 -  

8  Conclusion .................................................................................................................... - 57 -  

8.1  Thesis limitations .................................................................................................... - 58 -  

8.2  Future research ....................................................................................................... - 59 -
1 Introduction

For libraries that hold and preserve collections of archive material such as rare books, manuscripts, or photographs, managing these collections and making them available to patrons is of particular interest for the institution. In many cases, much of this material is of great potential research value. Efforts to digitize special collections provide the library with the chance to enhance and preserve the cultural heritage value these materials hold. Despite the benefits though, the time and effort needed for the library to register, digitize, and publish these resources, including enriching them for scholarly purposes, requires diverse skills, appropriate technology, and usable infrastructure to complete such a task.

Digitization, as a critical transmission activity, is not new to library initiatives or scholarly endeavours. Even so, projects wishing to exploit social web technologies have been hampered by efforts that, either, have not taken full advantage of the potential resources available, or else have been bound by traditional practices not readily transferrable or transmutable in the digital medium. In some cases, this has been because of a lack of understanding of how new models realize long-held conceptual beliefs. In others, practical application has been shown to be problematic due to the limitations of the tools available resulting in discrepancies between editorial aspiration and the ultimate achievement. As obvious as this may seem, awareness or assumptions about foundational principles may affect theoretical and practical decisions as well as drive performance from a technical standpoint. When undertaking digitization projects, it is beneficial for libraries and scholars to work together. Addressing both technical features as well as theoretical aspects can achieve best practice through collaborative interdisciplinary expertise.

The opportunity for more considered analysis of tool functionality and editorial practices has arisen by way of a project to digitize and transcribe J. A. Nensén’s records. The core project is being undertaken by Umeå University, but is part of a wider endeavour to make cultural heritage materials held by various Swedish libraries and institutions available and accessible through a common and collaborative platform developed by Uppsala University Library. The goal of this platform, Alvin (http://www.alvin-portal.org/alvin/), is not only to register and provide records of chosen materials, but to add value to them through scholarly research and publication of selected documents as facsimiles, transcriptions, translations, or digital scholarly editions.

By looking at how the group working with the Nensén records is performing transcription, this thesis will identify appropriate transcription tools for the project that fit such work. In the same manner, through this group’s exemplary activities, a study could potentially support and provide resources for other groups with similar aims. A description of the project will provide an overview and set it within a wider context: both as an individual transcription project and one in line with institutional endeavours.

This thesis first presents a description of the Nensén transcription project alongside the objectives and research questions of interest to this thesis. In Chapter 2, Background and Problem Description, a short description of Uppsala’s Alvin database places the Nensén project in context with wider library initiatives. Chapter 3, Conceptual Background and Critical Context defines transcription and provides the theoretical
model on traditional editing and transcription out of which this thesis builds its perspective for digital transmission efforts. Chapter 4, the Literature Review explores transcription via digital tools and in actual practice. Previous studies that are useful to this thesis are looked at. Methodology, Chapter 5, outlines the structure of this thesis’s approach to data collection and analysis of the results. Chapter 6 presents the results of the group interview and the tool review. First, the Nensén transcription model is shown. A tool selection process is then given from the results of the interview and the appropriate tools are identified. Lastly, the tool review is made up of those tools found to be most appropriate for the Nensén transcription model. Here, the tools are reviewed by how they enact transcription based on the model discovered earlier in Chapter 6 as well as traditional and theoretical editorial assumptions presented in Chapter 3. Finally a discussion on the benefits and problems discovered in the chosen tools is given followed by the conclusions of the thesis (Chapters 7 and 8 respectively). Limitations and future research are also identified.

1.1 “The Inquisitive Vicar”- digitization of J. A. Nensén’s records

The entire project consists of various parties performing different tasks in the process. The participants that make up the group will undertake everything from digitization to transcription, language study, and contextualization of the Nensén papers as this will meet their goals. However, this thesis intends to focus solely on the transcription stage of this project, and therefore, only those performing transcription.

The transcription group consists of two scholars at Umeå University in the Department of Literature and Nordic Languages. They will manually transcribe the source materials from scanned images over a period of two years (2015-2016). Previous digitization has been performed by personnel at the institutions holding the documents, and further work is intended as separate stages of the project requiring specialist expertise and resources.

1.1.1 The collection

According to the grant application written by the Nensén editorial group to Riksbankens Jubileumsfond (S. Karlgren, personal communication March 18, 2015) the collection itself is comprised of over 2000 pages of source material held between Uppsala University Library, Folkrörelsearkivet (Västerbotten’s Popular Movement Archive) in Umeå, and private holdings. Approximately 85% of the papers are written in Swedish, Norwegian, and Sami with the remaining 325 pages comprising Finnish, Russian, Latin, Livonian, and Polish. Roughly about 390 pages have been previously transcribed for research but not to the desired standards and do not do justice to the collection as a whole. Therefore, more comprehensive and systematic transcription must be put into effect.

The papers by Lapland priest J.A. Nensén (1791-1881) were written between the years 1813-1881. They cover a diverse range of topics including zoology, dialects, place names, pharmaceuticals, costumes, games, and specifically Swedish and Sami ways of life according to customs and traditions (S. Karlgren, personal communication March 18, 2015). The papers include extracts from parish registers as well as from books and dictionaries in various languages. The Nensén group’s grant application also points out
that the material is of international research value because the dossier specifically
denotes the North Scandinavian region and gives voice to many often neglected groups,
such as women, the poor, and the Sami. This is of interest to cultural geography and
gender studies, as well as linguistic, historical and ethnological studies.

1.1.2 Aim of transcribing the Nensén records

The outcome of the entire project, as presented to Riksbankens Jubileumsfond (S.
Karlgren, personal communication March 18, 2015), will be the full digitized corpus of
facsimile images and transcriptions of J.A. Nensén’s records made available and
managed through Uppsala University Library’s database Alvin in addition to being
published in parallel through Forskningsarkivet (The Research Archives) at Umeå
University Library; Silvermuseet (The Silver Museum); and Gaaltije (The South-Sami
Culture Centre). However, the main goal of the Nensén group is to make the materials
available for research (see section 6.1.1).

The transcriptions will be to philological standards and abbreviations will be expanded
(S. Karlgren, personal communication March 18, 2015). The grant application also
states that contextualizing introductions for the Swedish, Norwegian, and Sami
materials will be incorporated with the transcriptions at a later date. The documents in
other languages will be digitized and presented with the main collection, but not
transcribed at this stage. Keywords and metadata will also be added to aid in searching.

Transcription will enable the dossier to be more accessible for several purposes. First, it
will make the manuscripts easily readable due to the opaqueness of the text. Secondly,
an available corpus will be advantageous for researchers wishing to use the manuscripts.
This can also draw greater awareness to them for potential scholars and future research.

1.1.3 Practical considerations

Practical considerations for the digitization and presentation of the records were also
laid out in the application to Riksbankens Jubileumsfond (S. Karlgren, personal
communication March 18, 2015). The, approximately, 500 pages made up of direct
excerpts will not be transcribed. Instead, scanned images at the suggested resolution of
400dpi will be provided. Only the Swedish, Norwegian, and Sami materials will be
transcribed. The majority of the corpus (ca. 1250 pages) will go beyond the facsimiles
to afford transcription and indexing.

The bulk of the transcription will be undertaken by two researchers with palaeographic
expertise, and this activity, alongside incorporation into Alvin, will make up more than
half of the time-scale of the project. Each page is estimated to take 2 hours for Swedish
and Norwegian transcription and proofing and 5 hours for the Sami pages (S. Karlgren,
personal communication March 18, 2015). Besides copying the text, transcription will
consist of expanded abbreviations and documentation of page and line breaks. It will
follow a transcription model designed by the editors (see section 6.1.4). Metadata will
include identifying the informant’s names, gender and ethnicity, geographic location,
language used, and archival order. Biographical information and contextualization will
be added about the informants as well as persons of importance mentioned in the texts.
More collaborative work will be undertaken with Uppsala University Library after transcription to include data-technical corrections and implementation into their digital cultural heritage database *Alvin*.

### 1.2 Aims of the research

This thesis explores transcription and the tools available to editors. It sought to understand how the practical activities of the Nensén group can help in choosing an appropriate digital transcription tool. The objects of study are the tools themselves as well as the editorial workflow. By placing the available tools and the work application alongside theories of text, this thesis hoped to accomplish several overall goals:

First, add to the research on digital tools by understanding how transcription is performed through them.

Second, assess the transcription tools in critical digitization and build a selection model to enable other editors to choose appropriate ones for their projects.

Third, by placing it in a library context, this thesis hoped to support library research and bibliographic work as well as emphasizing critical digitization skills and knowledge as a library activity so that libraries are better able to support projects within their institutions and digital humanities as a scholarly discipline.

### 1.3 Objectives of the thesis

This thesis had three objectives: First, to observe the practices of the transcription group in order to understand and build a picture of their transcription model. In this case, the thesis looked at practice through a group interview. It tried to discover behaviours of the editors and how they approached transcription.

The second was to find available transcription tools. It identified these by conducting a selection and survey of the tools based on a literature review to find the features or functions the tools possess. This supported the first objective in understanding tool functionality via practice i.e. how they are problematic or beneficial.

And finally, this thesis reviewed those tools that seem most appropriate based on the editorial practices and needs of the Nensén transcription. A tool review was useful in presenting the most appropriate tools in light of the group’s transcription model. This occurred as an analysis of how they enact transcription. The benefits and limitations revealed enabled a more thorough understanding of critical digital transmission and assisted in building a more comprehensive picture of what is currently available for transcribers working with similar projects.
1.4 Research Questions

The qualitative and exploratory nature of this study aimed to answer several sets of questions in regards to transcription. The questions of interest were:

1. What is the group’s transcription model? How do they transcribe the manuscripts? What is problematic or desirable for them?
2. What are the transcription tools available? What characteristics or features do they have that can be used to review them? How should tool selection be approached?
3. How do the tools facilitate transcription? How do they support or hinder the transcription practices undertaken by the group? Do tool features meet the needs of the editors?

A group interview answers the first research question i.e. the group’s transcription model, and how they transcribe the manuscripts including what is problematic or desirable. The tool review answers the second question regarding the available tools, their characteristics, and how selection occurred. The final research question is then addressed. This tried to answer how tool functionality meets the needs of the editors by looking at how practice lines up with what the tool’s limit or make possible.

1.5 Limitations

Due to time constraints and the complexity of applying theoretical concepts to practical applications, this thesis needed to limit the number of tools being reviewed and could not anticipate full exploration of traditional editorial practices through them.

Tool selection was also dependent on many factors (see section 6.3). Goals may be different between projects or collaborators, and therefore, the chosen tools reviewed were limited by the users’ skills and the needs of the specific editorial group. In this way, tool specifications in these aspects guided selection. This thesis reviewed the tools based on goals of the Nensén project and the transcribers. However, in this approach, it is hoped that a deeper understanding of digital transmission and the tools available can be observed which will be helpful to other groups wishing to implement transcription tools in their projects.

Another limitation that affected the potential reach of this project was that the transcription group working with the manuscripts had only recently begun transcription. Therefore, the ability to see actual practices or completed transcription was determined by the editorial workflow and schedule of the transcription project as well as this thesis’s time scale.
2 Background and Problem Description

This study takes the stance that digitization projects are useful, desirable, and valuable in the field of library and information science, as well as disciplines in the digital humanities, for the explicit purpose that furthering the understanding of textuality in the digital world only aids cultural heritage projects that stem from this knowledge. Libraries can greatly enhance knowledge and practice in these fields through participation with, and research in, scholarly editing. This is one more way digital libraries are preserving and creating new uses for cultural heritage materials.

The project to critically digitize J. A. Nénsen’s records is in the position to benefit from research of this kind. By gaining a better understanding of actual transcription practice, digital tool functionality can be better evaluated. Setting practical application and a review of the transcription tools alongside theories of transcription and scholarly editing, this thesis will not only contribute to theoretical discussions of textuality and critical digitization, but it will also provide a framework for other groups involved in transcription projects for the collaborative platform Alvin currently under construction by Uppsala University Library.

The next section will give an overview of this project from a report provided by Krister Östlund (Personal communication, October 24, 2014).

2.1 Uppsala University Library’s digital archive and database

Uppsala University Library has developed a repository (ArkA-D) for the registration and presentation of digital collections. It aims to digitize the library’s personal archives and manuscripts. As a database tool, ArkA-D is built into the digital platform Alvin, or Archive and Library Virtual Image Network (http://www.alvin-portal.org/alvin/), which has been constructed with the intention of forming a partnership between multiple libraries to cut development and maintenance costs so that the focus remains on the digitization, registration, and publishing of these collections. The project intends to facilitate collaborative efforts for the digitization of unique cultural heritage material between Uppsala University Library, Lund University Library, Gothenburg University Library, and Linköping City Library.

As a portal for cultural heritage collections, it is hoped that Alvin will eventually contain the various special collections housed at these university and institutional libraries. Several external projects which are good candidates for registration and facilitation through Alvin have been identified and are being undertaken by separate groups who have expressed desires regarding digitization and electronic publishing. One is the group working to digitize and transcribe Lapland priest J.A. Nensén’s records.

Alvin will meet long-term preservation goals and ensure secure distribution for cultural heritage materials. It will enable individual groups to work with unique special collections while providing a usable and straightforward platform for a variety of projects. Therefore, technical issues and software will be addressed from the library’s end instead of by individuals.
The database behind *Alvin*, ArkA-D, is comprised of a four resource-level process (thus termed A-D) that aims to register and digitize the extensive archive collections found within the joint institutional libraries opening them up for access and use. The structure of this database is such that each level will enable the digitization process of the documents from a simple archival description through to a fully-fledged digital edition complete with comments and transcription as desired. The ultimate aim of ArkA-D seeks to facilitate the publication of these materials as useful scholarly resources. It is the final stage of ArkA-D which is of interest to the Nensén group and this thesis, as it is concerned with editorial practices.

Up to module C in ArkA-D, the work will be completed by the associated libraries regarding the documentation, digitization, and bibliographical activities of the archival collections. In this way, the project can be performed to international standards and quality controlled by the institutions housing them. Module A is the registration which includes basic documentation of the material along with its corresponding archive post. Module B includes digitization and image capture, connected with the third module, module C, where the appropriate metadata and library catalogue post is added to the facsimile.

The final stage, module D, assumes that the collection material is of value for publishing with supplemental information which can be likened to a full-text publication of the work. In some cases, if the work is of substantial research value, transcription can facilitate machine-readable, full-text searching of handwritten materials or even linguistic processing. Translation is also a possibility since the materials cover a range of languages. Module D has been designed with the idea of implementing a more open interface that allows editors to generate content fitting their needs and the unique characteristics of the collection material they are working with.

Since this stage will facilitate such various practices for different groups, it has been the library’s consideration to utilize external tools for such activities as transcription, translation, OCR, and annotation instead of building them into the interface itself. This will benefit the library by taking advantage of tools already in use and will save them time and money. Ideally, the tools should be open source and fit international standards for interoperability as well as drawing on the experience of other national and international projects which have used them. Since several of the projects have named transcription from facsimile images as a desired outcome, the library has begun looking at what tools are available. Currently, *T-Pen* is one tool under consideration, but the Nensén group has expressed a desire for a more comprehensive review of those available since their time-scale and skill-sets do not align with those of *Alvin* (see sections 6.1.1 and 6.1.3).

### 2.2 National and international perspective

Many libraries today are building and developing their digital collections. It is also more common for libraries to collaborate due to the expertise, diverse tasks, and cost needed to build and maintain these digital libraries. As a Swedish initiative, *Alvin* is in line with other European activities to uphold compatibility through standards and quality control. The intent to utilize a common platform for digital collections will make it possible for
Uppsala and the joint libraries to take advantage of the variety of competence available between these partners with the potential for external collaboration in the future.

Scandinavian efforts to critically digitize collections have been mainly collaborative, described more thoroughly by Dahlström and Ore (2013), such as the *Samnordisk runtextdatabas* (Scandinavian Runic-text Database) also through Uppsala University, and *Menota* (Medieval Nordic text archive). Many of the projects in these databases centre on the creation of digital scholarly editions that are of national and international importance as cultural heritage. Examples given by Dahlström and Ore (2013) include Almqvist in Sweden, Kierkegaard in Denmark, Ibsen in Norway, and Topelius in Finland.

Nationally, it is the aim of *Alvin* to centralize the digitization of cultural heritage material belonging to the participating Swedish libraries coupled with tools for cataloguing, archiving, and disseminating its content for a wider audience. The database and platform are currently in the early stages of development and construction having only been live since December 2014. *Alvin* has been financed by Riksbankens Jubileumsfond (The Swedish Foundation for Humanities and Social Sciences) and the entire project is financed by Kungliga Biblioteket (The National Library of Sweden), which in 2011, established the LUPP project with the intention of digitizing the collections of Linköping City Library and Uppsala University Library, among others, to offer users an expanded range of online resources (Andersson, Cullhed, von Wachenfeldt, & Östlund 2011) through a central platform with the tools to aid the digitization process.

Likewise, the project to digitize and transcribe J.A. Nensén’s records has also been financed by Riksbankens Jubileumsfond (Edlund 2014) having received a grant for two years (2015-2016) to make the folklore of Northern Scandinavia recorded by the Lapland priest available and accessible to researchers and the public alike.

More generally, the application of technologies and online publishing has brought new methods and possibilities to institutions wishing to digitize materials as well as affording researchers the electronic tools to edit and publish scholarly texts online. One example is University College London’s (UCL) *Transcribe Bentham* project (http://blogs.ucl.ac.uk/transcribe-bentham/) which has generated full-text transcriptions of a large portion of Jeremy Bentham’s personal manuscripts via crowdsourcing by amateur enthusiasts through an interface exploiting social media tools (see section 4.3 for more discussion on the *Transcription Desk*). This project has seen editorial practices change through new methods of transcription and digital tools. Such inventive modes of scholarly work are of value for prospective projects with similar goals and should be explored further.

This thesis sees that *Alvin* has the potential to house material that critical digitization practices can make available and accessible. Aimed to incorporate a type of social editorial practice via collaborative efforts in the digitization process, ArkA-D can provide the partnered libraries with the chance to develop their collections through digitization projects on an in-depth scale. Investigating one group’s actual methods will enable the appropriate tools to be identified for future efforts as well as deepening scholarly understanding of transcription through their application.
2.3 Problem Description

In initial communication with Uppsala University Library, it became apparent that several projects wished to transcribe their manuscripts alongside the digital facsimiles that would be available in Alvin. Uppsala had considered T-Pen as a possible tool to include in the database’s interface because it supported advanced transcription and collaborative projects (K. Östlund personal communication, October 24, 2014; P. Cullhed, personal communication, January 22, 2015) but had not come far enough in the implementation where a tool could be integrated into Alvin’s platform.

In further discussion with the Nensén group, who had earlier expressed interest in Alvin as a platform for their project (P. Cullhed, personal communication, January 22, 2015), they articulated their lack of knowledge about available transcription tools or awareness of any reviews of tools as they were not responsible for the technical decisions for the project (S. Karlgren personal communication, March 4, 2014). This is due in part to the lack of current research and studies in the field of digital transcription tools, and secondly, because the initial intent of the project was that the transcriptions would be performed via a tool provided by Uppsala and then uploaded to the collaborative portal Alvin.

However, since both Alvin and the Nensén project’s timelines began at different stages, a tool had not been decided on or provided, and the group have been left to make the decision themselves. This resulted in the editors beginning transcription as a separate effort with the intention of integrating the content into Alvin at a later date including opting for the most convenient solution presently available (see section 6.1.3).

Tool selection must consider issues of tool integration and performance, in addition to whether or not they fit each collaborating party. However, a review of each of the appropriate tools’ functionality from the group’s perspective is beyond the scope of this thesis. It is too timely and impractical to perform transcriptions of the documents for each tool. Therefore, this thesis will begin a selection of this kind by observing the actual transcription being undertaken by the group via their chosen tool, Microsoft Word. Focusing on what their practices bring to light, as well as their goals for the digital collection, this thesis hopes to place itself within the field of digital critical transmission to more fully examine how current tools could meet their needs.

In regards to this project, tool consideration must take into account usability and simplicity (as defined in section 6.3) as the actors engaged in the project are concerned with transcription only and not instrument design or implementation. On the other hand, the tools must also facilitate professional and suitable functions of scholarly editing. The problem here is balancing practice and goals with the tools available. One question could be: Are the existing tools appropriate for the tasks the editors are performing, and do they fit the outcome the project desires for the material? Since limited guidance in the form of previous research exists, this thesis seeks to fill this gap to some degree.
3 Conceptual Background and Critical Context

Transcription in the digital medium has an established tie with traditional transcription and editorial practices. Critical digitization is one area through which transcription has traversed this technological divide. It utilizes transcription by making the digital texts searchable and usable. Instead of simply providing a facsimile image of a manuscript, many of these texts are being keyed out or encoded to make them machine-readable and able to be manipulated for further use. Critical transmission has bridged the gap between print and digital medium, but despite novel advantages in the digital realm, transcription is still very much performed in traditional ways.

The next chapter, the Literature Review, will discuss the digital tools developed to perform transcription and how they have been used. First, though, it will be necessary to define transcription and look more closely at how editorial decisions are made that ultimately affect the final transcription. This will present a clearer picture of how traditional transcription understands a document and is able to represent it. Setting this against critical digitization can show where gaps occur between traditional practices and new media. This chapter will then go on to explore the theoretical assumptions present in transcriptions and how digital tools propagate or resolve these.

3.1 Definition of transcription

This thesis will take as its working definition of transcription from Vander Meulen and Tanselle (1999, p.201) who declare that a transcription is “the effort to report, in so far as typography allows, precisely what the textual inscription of a manuscript consists of”. It will begin from here because it will view transcription as two entities (i.e. transcription as task and as artefact) which are inevitably intertwined. This becomes pertinent when understanding the Nensén group as transcribing the source document to create a transcription that can stand (arguably) in place of the original.

Huitfeldt and Sperberg-McQueen (2008, p. 296) make the distinction between transcription as an act, as a product, and as relationship between documents i.e. one document is a transcription of another. This implies a certain derivation, or more specifically, some level of similarity between the exemplar and the transcription. That is, text can be seen as metaphoric, and hence, its graphemes or symbols can be interpreted or mapped to another with the same meaning to form a relationship between the two documents. Like the principal document, the physical mark is not transferable, but is in some way representative of the meaning it conveys depending on the agent chosen to identify it.

The definition of transcription is not confined by the editorial prerogative in deciding the level of transcription e.g. to use orthographic and normalized spellings or to represent precisely at the graphemic level, that is, the smallest unit of a writing system. These choices inevitably assume an interpretation, but are arguably of equal value where transcription is concerned. What this implies is that if detail or unique physical characteristics are distinguished between features or documents, the meaning is also distinct and must be treated as so unless otherwise stated. In order for a transcription to be successful for the purposes for which it was made, alterations must be selective and
defined for systematic and consistent transcription to occur. Therefore, transcription is not an established and defined act, but exists on a scale of degrees (Driscoll 2007, para. 1).

The Nensén group desires to provide the transcriptions as readable and available for research. They are historians, and as such, their editorial view will value that which will enable historical research. Content and language of the documents is of scholarly interest, and so linguistic features, names, dates, and descriptions of events will be ideally preserved.

3.2 An editorial task in the digital medium

Critical digitization, as defined by Dahlström (2014, p.16) is the faithful, digital representation of the source document. This is a critical transmission practice in line with such traditions as microfilming or manuscript copying by medieval scribes because it enacts bibliographic and editorial practices (Dahlström 2014, p.20). It is not simply a re-presenting of the text, but endows it with visual and graphical elements to support it such as markup. The resulting digital document might then be afforded additional metadata, indexing, descriptive encoding, paratexts, and bibliographical information. Additional scholarly materials may also be embedded or linked. “Critical digitization is qualitative in the sense that it concentrates on what is unique and contingent in the documents” (Dahlström 2014, p. 16), and is very much an editorial task.

Regarding the editor’s role in critical digitization enables a broader view in comprehending editorial work across boundaries and enhances the concept of transcription as a task being enacted on a document. Rosenberg (2007) takes the stance that the editor’s fundamental practice is unchanged in the digital medium; the editor still imparts his or her intellectual expertise on the documents no matter the form she or he takes. In this way, editorial practice can traverse medium to produce a transcription, and what is true in print form can be understood from the digital perspective.

What causes gaps in the digital medium, however, becomes more apparent in application (as will be seen from the group interview discussed fully in section 6.1). This is partly due to the fact that, as with traditional transcription, interpretation implies a perspective, (which inevitably is not objective), and partly due to the limits of each medium. Editorial intention may be hampered by what the chosen technology imparts or inhibits.

3.3 Theoretical assumptions

What is evident in scholarly discussion is that the digital medium has brought to light previous assumptions about the social and historical nature of text and the technology as well as editorial and critical methods that need to be addressed (Gabler 2010; McGann 2010; Shillingsburg 2010; and Siemens et al. 2012). Huitfeldt and Sperberg-McQueen (2008) point out two salient assumptions made apparent regarding transcriptions. First, they state that it is presumed that transcriptions are full transcriptions of everything in a document (p.307). This is detrimental for scholarly work being performed from the transcription as the latter is always incomplete in some way; it is an editorial
interpretation. Sperberg-McQueen (2009, p.31) makes this argument by stating that a text holds an “infinite set of facts”, any edition records a selection of the observable portions of these facts, and each edition provides some specific presentation of this selection. These acts (i.e. selection and presentation) are inevitably interpretive, which is in contrast to historical editorial objectives that sought to produce a “pure” text.\footnote{Textual theorists, such as Shillingsburg (1996), though, argue that the aim of discerning the “true” or “pure” text is a misguided approach considering the validity of all editorial beliefs and perspectives and current views of transcription value interpretation. For a discussion of the formal orientations held by editors see Shillingsburg, P. L. (1996). Scholarly Editing in the Computer Age: Theory and Practice. 3rd ed. Ann Arbor: The University of Michigan Press.} This editorial mentality (arguably no longer held) was carried over into the production of scholarly print editions, and perpetuated in digital editions as lamented by Robinson (2005), where the aim was an edition so definitive that it would never need editing again (Deegan & Sutherland 2009, p. 60).

Pierazzo (2011, p. 464) supports Huitfeldt and Sperberg-McQueen’s argument by stating: "no transcription, however accurate, will ever be able to represent entirely (i.e. faithfully) the source document", as the transcription must be systematic and consistent. It requires the expertise of a knowledgeable editor or group of editors to select and present the desired text. This is the essence of a textually-based scholarly edition. In any case, this interpretation has visible implications when it comes to digital transcription, including markup and text encoding, and how it is able to represent the editorial orientation.

The second assumption Huitfeldt and Sperberg-McQueen (2008, p. 308) highlight is the held belief that the text represented in a particular document, and thus the document itself, can be represented as a one-dimensional sequence of types. They argue this by pointing out that in practice, the document and text are structurally more complex than reflected in modern markup systems and that these different markup systems may have different notions of textual structure. This brings the focus back to the tools and their functionality as what remains questionable here are not the editors’ practices, but the abilities of the tools to allow for interpretive complexity: are they being designed by those who understand textual nuances?

A discussion about transcription cannot fail to mention a little more about markup as it correlates to assumptions about digital texts and is a type of transcription itself. Although the Nensén group will not be performing markup to a great extent, as digital texts, the group intend to make them accessible via a simplified encoding to guide later markup (see section 6.1.1). Therefore, a quick word about this will be useful.

As described above, a transcription can be defined as an artefact, but this artefact can be further distinguished as both the text itself and its encoding. Creating a machine-readable text may require markup in addition to copying the document’s contents. Buzzetti and McGann (2007) make the distinction between these two artefacts by seeing markup as coding the already bibliographically coded text, not the actual content. And here, transcription as an act is no different than in the print realm. Dahlström (2009, p.40) points out directly that markup itself is interpretive and enacts a theory of text just as the editor did before. Interpretation is always at the centre of any digital, critical transmission activity. A digital text (and subsequently its digital transcription) can be
understood as being made up of both the representation of the text itself and its structural content.

If the inescapable reality thus far is that transcription is a form of interpretation, whether desired or not, and that the representation via encoding is also reflected in the editor’s orientation, one conclusion is that the functionality of the markup system should support the editorial needs for the desired transcription. Therefore, gaining a better grasp of transcription practices can help elaborate on the nature of digital representation, preservation, and text encoding (Huitfeldt & Sperberg-McQueen 2008, p. 295). This thesis has tried to do just that by looking at the Nensén group’s transcription model and how it is translated using digital tools which, in turn, will be useful to evaluate tool functionality.

3.4 Editorial practice facilitated by digital tools

In general, editing requires specific skills. Editors involved in digitization projects may have goals that are separate from other editorial practices, and consequently, require specialized tools. Research has been limited regarding digital transcription practices and, ultimately the needs, of editors working with critical transmission tools and cultural heritage materials (see section 4.3).

Transcription as an editorial practice can be defined as representation (Unsworth 2000), a form of content capture, and is arguably the basis on which other scholarly activities are performed. That is, before discovery, enrichment, analysis, referencing, or even illustrating can occur for a printed and hand-written text, the representation of that content must be accomplished in some way. The tools required for working in these environments, which ideally must be easy and accessible for all (Robinson 2006, p. 11), must also fit the tasks and goals of the editors. Many current digital editions are still self-contained and their tools have not been fully realized (Robinson 2003, para. 6-7).

Critical digitization seeks to critically structure and organize primary sources and other heterogeneous data types for more coherent and centralized collections. Tools that facilitate these types of transcriptions encompass a range of features for editors from reproducing unique characters or symbols; automatic line parsing for linking facsimiles with the transcription; and correction or translation functions.

Many tools are built around standards that have been developed to support digital editing and transcribing. Organizations and initiatives, such as the TEI discussed at greater length by Price (2008) and Shreibman (2013), have spent great efforts creating and defining best practice guidelines for formal representation of documentation. Encoding a text through markup allows it to be machine readable as well as processed for a variety of uses. Tools that enable this and other standards, such as the Encoded Archival Description (EAD), and the Metadata Encoding and Transmission Standards (METS) make possible bibliographic, structural, and descriptive metadata across documents for their preservation and aggregation.

Research on transcription revolves around theory, but further work must be done to gain a practical understanding of exactly how it is performed through the tools available and how these tools limit, hinder, or make possible critical transmission.
4 Literature Review

This thesis is interested in critical transmission via manual transcription as it will focus on detailed transcription of a relatively small number of documents with limited or no automated workflow. It tried to answer questions pertaining to the group’s transcription model and how the available transcription tools can meet this. First, though, the current research surrounding transcription tools and practices needs to be explored.

4.1 Critical transmission

The very nature of critical digitization has been discussed as a bibliographic and editorial practice which implies an assumed orientation to make critical decisions about how editing will be accomplished and to what extent the text will be represented (section 3.2). New tools have opened up new possibilities for a variety of digital texts. Up until now, research regarding digital tools has focused mainly on scholarly editions to bring them into their new role by viewing them as community projects and environments facilitated by social media tools (Robinson 2003; McGann 2006; Mahony 2011; and Siemens et al. 2012). These studies are beneficial in that they allow critical transmission to be viewed in light of the tools used to create them. What is missing is a better grasp of how these tools enact editorial practices, specifically transcription. As Shillingsburg (2006, p.4) states, “there has been little effective development of a theory of electronic editing to support electronic editions, archives, or teaching tools”. More critical insight into digital editorial practices can help shape the current paradigm.

As argued by Siemens et al. (2012, p.446), most digital editions were created before current Web 2.0 technologies and, therefore, do not accurately reflect the possibilities for academic engagement or interaction. They are either print editions translated into a digital medium or else are hampered by traditional editorial methods trying to make use of new encoding models while retaining the expertise of skilled knowledge organization. In this way, research needs to extend the concept of critical transmission using these tools. However, most of the research surrounding digital tools and textual transmission reflects social media phenomena and popular trends. Annotation, folksonomy tagging, blogging, and text-analysis are indeed important and vital in creating dynamic, hyperlinked text and digital editions for a Semantic Web, but these applications deserve studies focused solely on each engagement.

4.2 Transcription

In library and information science, research on transcription and its related tools has seen a lot of attention in collaborative efforts (Causer, Tonra, & Wallace 2012; Walsh et al. 2014); spoken language transcription (Marsden et al. 2007; Garrard, Haigh, & de Jager 2011); and encoding principles (Driscoll 2000; Haugen 2004; Burnard, L., O’Keeffe, K. O., & Unsworth, J. 2007a).

---

2 See Siemens et al. (2012, p. 451-452) for a list of studies pertaining to various practices and social media tools.
Research on manual transcription of documents unfit for automated digitization, including OCR, is limited. Studies that consider materials similar to the Nensén documents tend to focus on the specific tools usually created to support these digitization projects, but they are useful as they provide a basis from which to depart. Two such tools are University College London’s Transcription Desk, and Ben Brumfield’s FromThePage further described in section 4.3 (see also sections 6.4.1 on FromThePage and 6.7 on markup approaches for both). These are not only well documented, but are valuable sources for identifying features and forming a foundation in which to compare their attributes. A third, T-Pen (http://t-pen.org/TPEN/), will be looked at because it has been identified as a possible tool for the Nensén project as well as other transcription projects identified for Uppsala’s Alvin platform (see section 6.4.4).

Since manual transcription of handwritten manuscripts, as opposed to automated transcription or OCR procedures, is vital to the Nensén materials for the production of searchable, encoded digital texts, there is a requisite for more research surrounding these practices and how the available tools designed for these objectives meet scholarly needs.

4.3 Transcription tools in research and practice

The Transcribe Bentham project (http://blogs.ucl.ac.uk/transcribe-bentham/) undertaken by the University College London (UCL) has been a major endeavour that has examined how editorial practices are changing in light of new technologies and methods (Causer, Tonra, & Wallace 2012). The project developed the Transcription Desk (http://www.transcribe-bentham.da.ulcc.ac.uk/td/Transcribe_Bentham) which was a tool specifically designed for the transcription of Jeremy Bentham’s personal handwritten manuscripts. The tool itself has been specifically designed to facilitate crowdsourcing transcription with TEI-compliant XML. This enables volunteers with little expertise to perform the task to a desired degree. The TEI subset, located in a toolbar (see section 6.7), makes the addition of markup possible by clicking a corresponding button. This feature seeks to automate part of the process of a task that requires many volunteers working over many years.

The Transcription Desk can stand as a valuable source in tool functionality and features when looking at the other tools available due to its extensive use in the Transcribe Bentham project and supported research. It is, however, not reviewed in this thesis. This is because it is built on a MediaWiki and therefore similar to Scripto, Wikisource and other plugin-based tools. Since these others are already discussed in depth, and further exploration of various MediaWikis is outside the scope of this thesis, the Transcription Desk’s TEI tool bar will be the most relevant to touch on. Instead, the Transcription Desk will better serve as an example of critical digital transmission practice. More specifically, how it has been used and how it has changed transcription in the digital medium will be more beneficial as a reference point. Limitations, benefits, and assumptions about text inherent in other tools are made evident and can be more easily assessed when seeing how practice has been affected by an already widely used tool.

As a collaborative tool, the Transcription Desk was designed for non-expert editors and transcribers. It was built with simplicity in mind, including a plain-text box into which users could type, and basic formatting functions (Causer, Tonra, & Wallace 2012, p.
One conclusion of the study by Causer, Tonra, and Wallace (2012) was that editing practice was changed. Transcription was originally performed in several formats, some denoting different layers of transcription such as format or short-hand. However, with the TEI subset, transcription could occur alongside for functional and meaningful identification of the words (pp. 123-124). The *Transcription Desk* also allowed easy conversion of the transcription to an XML file (p. 125). This along with the TEI encoding improves document accessibility and refined searching, thus allowing editors and other project actors to meet overall goals for the Bentham papers and the digital archive.

Another tool, *FromThePage*, was also designed for a specific project by the software engineer Ben Brumfield. It is now widely used in various other projects and strongly documented by Brumfield himself (Brumfield 2014b). Unlike the *Transcription Desk*, however, *FromThePage* does not use TEI markup. Instead, it uses Wikilinks subject identification to mark the text in plain text. This has several benefits most obviously avoiding the prerequisite of learning the TEI principles. Helpful to this study is the opportunity to look at transcription that does not require the TEI encoding principles, as it is not always useful (Lavagnino 2007), but still able to build documents that can be treated like databases for in-text searching and text processing (see section 6.7 for a discussion on markup).

*FromThePage* is described more fully in section 6.4.1 as part of the tool review, but it is worth noting that as a tool, it was also designed to be as simple and usable as possible for projects employing volunteer editors. Research regarding this tool is related to its functionality and limitations which can be useful to defining tool characteristics and comparing usability and performance between them.

### 4.4 Research methodologies

Research purporting reviews of transcription tools is limited and has focused mainly on automated audio transcription (Marsden et al. 2007), but this study can, nonetheless, be a guiding methodology. Here, the study was grounded in the assessment of actual practice. The tool’s functionality could be measured as to how well it performed the tasks specific editors required. Brumfield (2014b) and Pierazzo (2011) support this in their claims that tool choice should be made based on the goals of the project. Conducting a focus-group as performed in Marsden et al. (2007) would, therefore, be optimal in understanding editorial choices to guide selection of the tools to be surveyed. Since the decision of the tool will necessarily be affected by the goals of the editors, it would be useful to understand their transcription methods and the goals of the project. Marsden et al. (2007) suggest that semi-structured interviews are a way to assess user-needs followed by more in-depth interviews and oral histories to understand the work process. Here, the Nensén editors would be able to clarify a typical transcription task with concrete examples showing how transcription is performed and what elements present problems for them in the records.

Determining how the tools are to be evaluated can be established based on needed transcription tasks of the editors (Driscoll 2007; Pierazzo 2011) and will form the basis for a list of tool features to be compared and contrasted. These include such considerations as how the editor will approach what Driscoll (2007, para. 1) describes
as documentary substantives (i.e. actual words of the text) and accidentals (i.e. surface features, spelling, punctuation, word division). Pierazzo (2011, p. 468) sees a contrast between what she calls graphic evidence and meta-information making the distinction between physical components that can be represented in almost any system and semantic and grammatical functions which require encoding such as with XML.

The next chapter will outline how this thesis accomplished its objectives in more detail.
5 Methodology

This thesis proposed several research questions that intend to investigate traditional transcription practices translated into a digital environment. In so doing, the thesis explored the transcription of J. A. Nensén’s records to understand critical transmission as an editorial endeavour. A thorough account of the design, data collection, and data analysis of the study will ensure the validity and reliability of the research process and corresponding data (Denscomb 2009, p.155). It will also justify the findings and subsequent results.

5.1 Research design

The exploratory nature of the research problems addressed by this thesis necessitated an approach that could provide qualitative data gathered by multiple means along with the ability to intensely study the practices of a specific group. A case study, as discussed by Choemprayong and Wildemuth (2009), was the most appropriate strategy for a research goal of this kind since the research questions posed sought to address the constructs of phenomena and describe them. This method requires a loosely structured process with a flexible design (Wildemuth 2009, p. 15). However, this thesis was not able to monitor the group’s practices over a long period or perform follow-up interviews. Therefore, the chosen method of data collection was an in-depth semi-structured group interview.

The research methodology was designed out of the objectives of this thesis:

1. Create a picture of the editors’ transcription model
2. Create a selection process to identify appropriate tools
3. Review transcription tools that fit the group’s practice

The latter goals naturally stemmed from the results of the first so that these could be performed in succession.

5.2 Data collection methods

The relevant data was collected through various means to support qualitative research. It used a group interview and a literature review to gather appropriate data to answer the research questions regarding the group’s transcription model, available transcription tools, and how tool features enacted or inhibited the desired model laid out in the research questions (section 1.4). An approach able to address each question is outlined by the following structure:

1. Explore the practices of the transcribers to identify what was problematic or desirable in their current model.
2. Identify the transcription tools and their features and select the appropriate tools to be reviewed and,
3. Investigate the tools’ functionality to see how they fit the Nensén group’s needs and practices.
To explore transcription practices, an interview was undertaken using a semi-structured format. This was complemented with follow-up correspondence via email. A group interview was chosen for several reasons not least of all its advantages in assessing activities and gathering qualitative data.

First, the number of editors transcribing the Nensén records was small enough for a group discussion to be beneficial. This permitted more in-depth information to be collected based on personal experience, and a conversation-like format was more likely to take place and would permit interaction and dialogue between the participants actively working together. The group size was also favourable as the entire population’s views could be assessed without the need for probability sampling.

Second, a group interview would validate the inductive nature of qualitative content analysis to support the wider context of transcription practice. This was desirable because, as Sperberg-McQueen (2009, p. 31) states, transcription has an “infinite set of facts” and consequently, preparing a complete list of features or characteristics beforehand would not be possible. The semi-structured format also facilitated the discovery of missed or hidden properties.

Identification of tools and their functionality was performed first through a literature review of surrounding documentation and then a more in-depth literature review and tool exploration for tool features. This is described more fully in section 5.5.

5.3 Participants

The relevant population for the interview consisted of only two editors directly involved with transcription of the Nensén records. The group had been found through previous discussion with Uppsala University about current heritage projects interested in the Alvin database (P. Cullhed, personal communication, January 22, 2015). The Nensén project leader, Lars-Erik Edlund, though not acting as a transcriber, was of value to the group interview as he was directly involved in building the transcription principles and theoretical model. However, he was unavailable for the interview. Data in this regard was therefore collected through emails and a phone interview to compliment and expand the information provided by the editors themselves.

5.4 Design of the group interview

An interview guide was prepared (see Appendix I) in order to gather information on how the group would approach their transcription. Questions were developed based on previous research (section 4.3) surrounding transcription tools and the projects they supported. Here, it was observed that tool selection was highly dependent on the definition of transcription employed by the editors and their goals. Therefore, it was considered useful to allow for a semi-structured composition of closed and open-ended questions aimed at collecting focused data on those issues that were most relevant to the Nensén group. The questions also left room for open discussion to reveal other areas of significance.
The questions were divided into four main categories: Goals of the project and the editor’s themselves; practical issues concerning technology and workflow; information about the source material; and transcription practices and decisions. Although some of these questions overlapped, the structure allowed for an organized and consistent flow to the interview.

The questions pertaining to transcription were designed to develop a detailed picture of the group’s actual transcription activity. This section was subsequently divided into two sections to assess how the documents would be interpreted. These involved accidentals (i.e. surface features and graphic evidence) which can be seen in the edited document and meta-information (i.e. semantics and grammatical functions) which are not apparent in the transcription and must be recorded using an annotation system like XML.

The final section of questions for the interview focused on specific and unique characteristics about the Nensén records.

The interview took place on April 16, 2015 at Umeå University. A café on the university campus was chosen to allow the interview an informal setting and discussion to be more relaxed. It was conducted by the author acting as moderator. The entire interview lasted three hours, including a break for coffee, and was audio recorded while notes were taken simultaneously. This way, less time was spent by the author in pausing between comments and thoughts to write. Additionally, the group interview permitted visualisation of a facsimile placed side-by-side its transcription and occurred in the editors’ office after the initial discussion. This validated the group’s comments enabling them to demonstrate practical issues and the solutions they had arrived at through physical examples and illustrations in Microsoft Word.

The interview was transcribed in part, that is, it was edited to remove pauses, redundant vocalizations, and the rephrasing of questions and answers to improve the flow. The interview occurred in two languages (i.e. Swedish and English) and the transcript followed this since some terms were more descriptive in one of the languages. Since the group discussion included many social and non-verbal elements, the transcription was unable to capture everything. Where something such as a demonstration or visualization occurred that was of importance to the question, the transcript used brackets to explain what was being done but was more typically documented in the notes for reference. For the most part, it was thought contextually unnecessary to indicate more than basic agreements, disagreements, and opinions.

Communication with the editors through email was used as follow-up to support and compliment the data collected from the interview.

5.5 Identifying transcription tools

To answer the questions pertaining to what transcription tools are available, and their characteristics useful for a review, identification of current tools was first required. This was done through a literature review of the software and corresponding documents. As virtual documents (e.g. websites, software manuals, blogs), it was essential to ensure the authenticity, credibility, and timeliness of the documentation (Bryman 2008, p. 525). Therefore, close attention was paid to the source websites. The tools selected were from
reputable authorities and developers. They could also be evaluated through their integration and utilization by other credible institutions and projects. The tools were identified through the following websites: The DiRT Directory (http://dirdirectory.org/); the Text Encoding Initiative website (http://www.tei-c.org/Tools/); the Digital Humanities Now website (http://digitalhumanitiesnow.org/); digitization.eu (http://www.digitisation.eu/); and Ben Brumfield’s blog Collaborative Manuscript Transcription (http://manuscripttranscription.blogspot.se/).

Transcription features and functionality were identified in the following literature and websites: Driscoll 2007; Pierazzo 2011; Brumfield 2012b and 2014b; and The University College London’s (UCL) Transcribe Bentham project’s Transcription Desk (http://www.transcribe-bentham.da.ulcc.ac.uk/td/Transcribe_Bentham). These features then helped construct a picture of the group’s transcription model: the editorial choices they made that could be expressed via tool functionality.

Twenty-eight text tools and their properties were identified (see Appendix II). It should be mentioned that Brumfield’s (2012b) Crowdsourced Transcription Tool List was the inspiration for the design of this thesis’s list, although many of the tools in his list were designed for specific projects and have been subsequently left out as they are not applicable or readily available from the developers. Also, since more features were wanted for evaluation, and discovered in the literature review, it was deemed necessary to expand the categories in the tool list to aid selection process. Since the tools were to be based on their suitability and functionality, a more detailed grid of features was developed.

It was discovered that due to the thesis scope, not all the tools relevant for the Nensén project could be reviewed. Refining tool selection necessitated making decisions based on how appropriate they were for the Nensén group. Therefore, a method for delimiting was required to select those tools that could best fit the Nensén group’s transcription practices. Using the themes and keywords from the interview, the wide variety of relevant tools could be narrowed to just those that were more representative of the editor’s practices.

5.6 Methods of data analysis

Qualitative content analysis was used as the method of data analysis for the interview. This detected themes and discovered common or unique properties (Zhang & Wildemuth 2009 p. 309) that could be used to construct and understand the Nensén group’s transcription model. Qualitative content analysis was performed by coding the interview transcript. Bryman (2008, pp. 550-552) outlines a method in which coding can be performed. This acted as a systematic, established authority to justify the results of the analysis. Bryman suggests, first, a read through of the transcript without taking notes. A few general thoughts can be recorded afterward. This is then followed by a rereading of the interview and making notes on significant remarks or observations. Reviewing these notes allows common words or phrases to be seen and their connections with the concepts in the literature. The notes can then be refined to codes and themes. Reoccurring concepts or ideas were noted in relation to the question sets (i.e. goals, material, technical considerations, and transcription) and then grouped into overarching categories which became the themes out of which analysis could occur.
It was observed that there was a lot of overlap between the question sets and the themes. For example, diplomatic was a concept expressed in relation to both the act of transcription (based heavily on the physical nature of the material set against what Word would allow them to do) and the editors’ goals for the project such as preservation and readability. The themes discovered were:

1. Context  
2. Diplomatic  
3. Readability  
4. Simplicity

It was observed that another theme occurred relatively often in conjunction with the question sets as well as in addition to the other themes. It was collaboration. As a theme though, collaboration, will not be fully explored in the scope of this thesis as the actual transcription will not be collaborative. Nonetheless, it was helpful in limiting the number of relevant tools able to be reviewed in scope of the thesis. Some tools that were seen as relevant were eLaborate, Digital Mappaemundi, and ediarum. Collaborative tools, however, are usually built for users with limited technical skills, as in the case of the Transcription Desk, and take advantage of other parties’ expertise. Since the Nensén project is such a collaborative endeavour (albeit not fully implemented), tools built to support and bolster these types of partnerships are more suitable. Therefore, the thesis was able to limit tool selection even further. Consequently, the tools were narrowed to four (FromThePage, Scripto, T-Pen, Wikisource) enabling a more thorough evaluation based on their appropriateness.
6 Results

This chapter will present empirical data to provide a detailed overview of the results of the group interview and tool identification. It will also consider functionality by dealing with how the group approached editorial decisions through the means they had available. However, a fuller account of this research question will be presented in the Discussion chapter. First, though, the results of the interview aim to answer the first set of research questions: What is the group’s transcription model? How do they transcribe the manuscripts? What is problematic or desirable for them?

6.1 Group interview

The group interview provided qualitative data in the form of an interview transcript and notes that were subsequently coded as discussed earlier in sections 5.1.3 and 5.1.5. This enabled a qualitative data analysis of the dense and complex nature of the information that used the previously determined themes to find patterns and gaps as well as discovering other relevant concepts that were not initially evident. The group interview gave a detailed picture of the transcription model developed by the Nensén project which was used to visualize their transcription theory and practice using Microsoft Word and then assess tool performance.

6.1.1 Goals

The first set of questions focused on the goals of the project and the transcribers. The interview questions differentiated between presentation, use, users, and markup. Discovering the goals of the group would help define how the final edition might be presented and used as well as on what aspects of the project the group will concentrate their efforts. This corresponds with the “purpose” of a transcript laid out by Brumfield (2013) as a fundamental factor for transcription tool selection.

The main goal of the Nensén project is making the materials available for researchers. Therefore, digitization, transcription, and translation are all equally important for such a task. Attention to transcription, though, is the aim of this thesis’s study. Transcription will make the documents accessible, namely, in making them available and readable: Available, because the material is currently held in different locations, and readable, because of the density and abbreviated nature of the text.

The consensus was that presentation should provide the facsimile image alongside the transcription with additional full-text searching and a pop-up window that would offer definitions or commentary when the cursor was held over a word or sentence. This would enable the user to view the image and text side-by-side for personal comparison. It would also place a level of engagement with the text in the user’s hands as she or he would be able to choose to view the text with the added scholarly research or not.

Parallel publication was also a goal of the transcriptions. The Nensén group discussed having their own website which would be built by Umeå University Library and made available through Forskningsarkivets (The Research Archives’) website. The other
institutions who were interested in disseminating the material were, of course, Uppsala’s *Alvin* database, Silvermuseet (The Silver Museum), and Gaaltije (the South Sami Culture Centre). Each of these institutions (including the transcribers themselves) had an interest in the outcome of the overall project. Collaboration was a key issue in the project as each of the institutions played a role.

However, it soon became clear in the dialogue that the goals spanned beyond presentation and use of the materials as the project’s objectives were seen differently from each participating perspective i.e. the individual transcribers; the Nensén project including future translators and researchers; and Uppsala University Library’s *Alvin* group. This was expressed by Karlgren (Personal communication, April 16, 2015):

[B]efore we started, we thought we could have pictures beside the text [...] and you could scroll them, and if you put the pointer here you would be pointed in the text. And that’s what we wanted when we started. But when we talked to *Alvin* again it turned out they just wanted the pictures for now and not the transcription. [I]t was more like *[Alvin]* wanted to preserve the originals for now, and then we can work with [the materials] later on.

The main discrepancy identified in this respect was between short-term and long-term goals, not so much differences between the groups’ ultimate aims. This stems from the fact that *Alvin* is currently in its early stages. This issue supports the overall need to understand the individual group’s process while the editors are under time constraints to complete transcription.

Therefore, coding sought to identify the goals of each of these separate groups in order to understand only those gaps that affected transcription. This is due to the fact that this thesis does not intend to look at the dynamics of collaboration. However, surprisingly enough, the goals were categorized in such a way that reveals similarities between each of the groups. This reinforced the need for collaboration, specifically, the transcribers’ need for the library to facilitate transcription by making the technical decisions and building the infrastructure as Umeå University Library will publish in parallel through another website. Since their goals are purely based on the transcription, the transcribers felt they did not have the expertise to make decisions regarding tool choice. One comment by Karlgren (Personal communication, April 16, 2015) illustrates beautifully their need and the potential needs of other groups contributing projects in *Alvin*:

The easiest would be if *Alvin* just chose the instrument we’re using […], that would be the best for me because what we are doing is just making this available and if they could just hand us the instruments to make it available that would be good as well. And there are so many projects in *Alvin*, so if you want it to be coherent it would be good.

The general goals of all the groups were thus identified as: Collection building, preservation, visibility, availability, and readability. The extent to which they are short-term or long-term depends on the specific group. A description of how transcription affects each of these is as follows:

1. Collection building- Scanning the documents to produce facsimile images will allow all the material to be gathered in one place since the documents are in
Uppsala, Umeå, and even a private collection. Transcription of these can build a thematic research collection of the material as scholarly works.

2. Preservation- Facsimiles will preserve the material by acting as surrogates. Transcription will preserve the documents in the same manner.

3. Visibility- Facsimiles and transcription provided online and through various institutional websites will make the material more visible to researchers and general users. It may also increase interest in the documents due to their being available.

4. Availability- As with visibility, digital copies will be more accessible to researchers or users wishing to take advantage of the material. Since there is already national interest in the sources, this will make the material available for future research.

5. Readability- Based on a user-perspective, transcription will make the documents readable which before was difficult due to the nature of the text. Full-text transcription and keywords will enable searching. This will also add value to the material as much of the abbreviations or terms are unfamiliar to modern readers. Translation will make the material accessible for those who wish to read the sections written in other languages.

Therefore the tools should be reviewed by how well they can enable these goals to the general standards. In this they need to consider how these goals are met and if they meet the transcription goals for the edition that the editors desire. Issues such as open source licensing might be up for debate if costs will be covered by Uppsala instead of the individual groups, while a tool that allows side-by-side views of the facsimile and transcription and/or translation might be more firmly set as a standard since all the actors are seeking readability from the user-perspective.

One goal of the Nensén project, and specifically the transcribers, was that it had been decided that the transcription should be as diplomatic as possible. This will be discussed more fully in section 6.6, but it is of importance in two ways. First, it supports the claim that their transcription model is necessary when reviewing the tools. In consequence, what the tools allow them to do will affect either the tool choice or the model must change to accommodate the tool. This will ultimately alter the edition they wish to present. Second, it generally applies to any group undertaking a similar task. That is, if Alvin decides to adopt one instrument for all the projects wishing to contribute, it must meet a standardized transcription practice acceptable to all contributors.

One transcription goal that is of importance in tool review but was not applicable to the Nensén group was markup. This was a very interesting issue as it was clear in the interview that markup via XML or the TEI scheme was desired and considered vital, but the transcribers were not the ones to do it. Also, there seemed to be a distinction between textual or linguistic markup (i.e. meta-information) that is performed via encoding and markup that is semantic as revealed by Karlgren when she says, “[markup] is what we are handing over. To the people working with the computer stuff! But we are kind of marking them up as we go because we mark them and comment in the side lines” (Personal communication, April 16, 2015). Although, Karlgren also
points out that they are doing a type of “prework” by adding keywords and commentary which will guide the future encoding. But here it seems to suggest that markup is a task apart from traditional transcription and editorial theory.

6.1.2 The material

The second section of interview questions focused on the material. Physical characteristics of the documents themselves have been shown to be important when considering transcription (Pierazzo 2011; Brumfield 2014b), and specifically, in how the editors will address certain issues with the material.

A general overview of the collection provided by Eckeryd (Personal communication, April 16, 2015), found that it is spread out over several locations. In one instance as loose leaf pages collected in two volumes held by Uppsala University Library as well as material later discovered and collected and stored loose in a box. Folkrörelsearkivet (Västerbotten’s Popular Movement Archive) possesses some donated material, while a diary is still in private ownership (L. E. Edlund, personal communication, April 27, 2015). Although the division of ownership means the materials reside with the respective holders, it does not so much affect the ability to publish the material nor make it freely available. Similarly, Riksbankens Jubileumsfond has funded the project and this cultural grant allows for the material to be made freely available. Likewise, the diary has been made available for scanning by the private owner as it is of value to the collection as a whole.

The source documents are compiled of handwritten text in different languages (see figure 1). The text is in a Latin script which is common for the period in which the records were made: 1818-1881 (R. Eckeryd & S. Karlgren, personal communication, April 16, 2015). The handwriting is neat and consistent, but the pages are fully covered with heavy abbreviations and insertions making reading difficult (see figure 2). Eckeryd describes how many of the documents are simply pages folded in half to create small booklets of four pages. It is known that there are folios missing as some pages begin in the middle of an entry as well as other folios identifying that they are specific pages out of a total number which are not present.

The content of the manuscripts, as Eckeryd and Karlgren (Personal communication, April 16, 2015) describe, is made up of records. This consist of descriptions, histories, and observations including names, dates, places, and other biographical and factual information as well as passages or words in dialect variants and other languages. Each entry is numbered and contains the information pertaining to the topic (see figure 2). In places it is clear that Nensén left blanks to compliment the subject with information, such as deaths or contextual references, some of which were never filled in. The source material, therefore, is mainly text with divisions such as headings and paragraphs. In many of the pages, he has divided the folio into columns or sections (see figure 1) by spaces, lines, or even text that runs vertically.
Figure 1 Folio 22 from J.A. Nensén’s records, provided by S. Karlgren, Umeå University

Figure 2 Detail of folio 19 from J. A. Nensén’s records, S. Karlgren, Umeå University
6.1.3 Practical issues

The third set of questions for the group interview sought to understand technical needs in a practical way and the general workflow of the group. This was because Brumfield (2013) identifies financial, technical, and organizational aspects as considerations in tool choice. Among these include such factors as system administration, budget, programming skills of staff, hosting considerations, and whether encoding is mandated in the workflow.

For the most part, it was clear from the goals of the Nensén group that simplicity was the main aim: “[The argument was] ‘håller det enkelt’, ‘keep it simple’” (S. Karlgren, personal communication, April 16, 2015). Also, as was described previously in section 6.1.1, the transcribers themselves were not responsible for the technical decisions nor had a decision been made regarding the tool since Alvin was initially focused on preservation. In regards to their own website, Forskningsarkivet (The Research Archive) would determine specifications, and Uppsala University Library would manage the decisions for Alvin.

Previously, Alvin identified T-Pen as a possible transcription tool (K. Östlund, personal communication October 24, 2014; P. Cullhed, personal communication, January 22, 2015) due to the desire to encode text with TEI standards. However, software integration was seen as problematic, and this was the reason it was dropped. These issues led the Nensén group to choose to transcribe the documents in Microsoft Word for simplicity’s sake (S. Karlgren, personal communication, April 16, 2015).

The project was designated to proceed under a two-year time scale. The majority of the documents were first scanned in Uppsala under standardized specifications. The workflow consisted of transcribing only the Swedish, Norwegian, and Sami material because of the transcribers’ expertise, as well as the goals of the project and the degree to which the funding covered. It was decided excerpt passages were not to be transcribed and provided only as facsimiles. Initially, transcription began with the Swedish material “to get a feel of how he writes certain letters because otherwise we would have made more mistakes with the Sami material” (R. Eckeryd, personal communication, April 16, 2015). It was decided to complete the Sami material first as it could then be sent for translation. Since much of the text was abbreviations, it was decided that these, and other symbols, would be expanded for readability. A list of abbreviations and symbols was created and continuously updated for reference.

6.1.4 Transcription model

The group’s transcription model was the aim of the fourth and fifth set of questions. These questions sought to identify specific transcription practices and unique issues surrounding the Nensén records. A section devoted to the encoding of the text in XML was developed. Already discussed as part of their goals, it became apparent that this was desired but not in the preliminary project design. Therefore, the interview adapted the questions to a more detail-oriented perspective of how the transcribers handled specific issues in practice using Microsoft Word and how this tool affected outcomes.
It was decided to approach the transcription model in a way that would allow a clearer comparison with the tools but still understand it in terms of the chosen tool, Microsoft Word. Therefore, table 1 was developed to enable a more systematic comparison to tool functionally discussed later and shown in table 2.

The transcription model is very specific to the Nensén group, and should be seen as such. The table is an attempt to visualise their choices, and when compared with the latter table, a means of revealing inconsistencies and assumptions in transcription across media since transcription cannot be simply understood in terms of exact reproduction (Huitfeldt and Sperberg-McQueen 2008).

Table 1 shows an overview of the theoretical concepts and how they were approached, but a more detailed description of each feature follows which will be useful in clarifying the transcription model presented in the table. The first half deals with glyph properties, the second part with topographic or surface features of the text.
Table 1 The Nensén project transcription model in Microsoft Word: feature decisions

<table>
<thead>
<tr>
<th>Accidental theme</th>
<th>Properties</th>
<th>Preserved as expressed?</th>
<th>Reason</th>
<th>Expressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orthography</td>
<td>Spelling variations</td>
<td>Yes</td>
<td>Diplomatic</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>No- if unclear</td>
<td>Readability</td>
<td>(see abbreviations)</td>
</tr>
<tr>
<td>Misspellings</td>
<td>Yes</td>
<td>Diplomatic</td>
<td>Correction in square brackets</td>
<td></td>
</tr>
<tr>
<td></td>
<td>If unclear</td>
<td>Readability</td>
<td>[= word]</td>
<td></td>
</tr>
<tr>
<td>Diacritics</td>
<td>Yes</td>
<td>Diplomatic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Writing Features</td>
<td>Capitalizations</td>
<td>Yes</td>
<td>Diplomatic</td>
<td></td>
</tr>
<tr>
<td>Punctuation</td>
<td>Yes</td>
<td>Diplomatic</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>No- if they are/ indicate abbreviations</td>
<td>Readability</td>
<td>Expanded with missing letters in italics</td>
<td></td>
</tr>
<tr>
<td>Symbols</td>
<td>No</td>
<td>Tool functionality, Readability</td>
<td>(see punctuation)</td>
<td></td>
</tr>
<tr>
<td>Superscripts</td>
<td>Yes- ordinal numbers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abbreviations</td>
<td>Known</td>
<td>No</td>
<td>Diplomatic, Readability</td>
<td>Expanded: missing letters in italics</td>
</tr>
<tr>
<td></td>
<td>Unsure</td>
<td>No</td>
<td>Diplomatic, Readability</td>
<td>Expanded: put in square brackets</td>
</tr>
<tr>
<td>Letter forms</td>
<td>Latin script</td>
<td>No</td>
<td>Tool functionality</td>
<td></td>
</tr>
<tr>
<td>Ligatures</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Letter variations</td>
<td>No</td>
<td>Tool functionality</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Surface features</td>
<td>Line breaks/ page breaks</td>
<td>Yes</td>
<td>Diplomatic</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Paragraphs</td>
<td>Yes</td>
<td>Diplomatic</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Columns/ lists</td>
<td>Yes</td>
<td>Diplomatic</td>
<td></td>
</tr>
<tr>
<td>Headers</td>
<td>Yes</td>
<td></td>
<td>Diplomatic</td>
<td></td>
</tr>
<tr>
<td></td>
<td>No- depending on text layout</td>
<td>Readability, Tool functionality, Context</td>
<td>Placed in single tilting marks: ‘between lines’, ‘vertical text’</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>Drawings/ excerpts</td>
<td>Yes</td>
<td>Diplomatic, Context</td>
<td>Facsimile</td>
</tr>
<tr>
<td></td>
<td>Document genesis</td>
<td>Yes- generally</td>
<td>Diplomatic</td>
<td>As an introduction</td>
</tr>
</tbody>
</table>
Orthographic elements include spelling, spelling variations, misspellings, and diacritics (i.e. glyphs added to letters). In the transcriptions, all spellings are kept to diplomatic ideals. For example, Nensén uses “f” with or instead of “v” and “w” as in the case of the words af [av = of] and elfwen [älven = river]. This is the case with misspellings too except where there is a big mistake or one that is hard to understand. In this instance, the spelling is retained but the word is indicated in square brackets with an equal sign. For example, Nensén sometimes spells names differently. These are regularized in brackets if it is difficult to read “so people know what he’s talking about” (S. Karlgren, personal communication, April 16, 2015).

Writing features such as capitalizations and punctuation are kept to create as diplomatic an edition as possible. Punctuation, however, is sometimes unclear. For example, Nensén uses a long under-scored mark [_ __] (see figure 2). As Karlgren (Personal communication, April 16, 2015) reflects:

[I]t is a kind of punctuation and what we decided is to keep it as a line, as a base line, because it is both a pause and punctuation. It is something in between [a hyphen and an underscore] so we keep it like he uses them. And sometimes you have a dot and then the line. That is an abbreviation and that also indicates that it is not a dot per se, it is something else.

The same is true for symbols which can stand as abbreviations, as many of the ones Nensén uses are similar to the ampersand [&]. As an example, Nensén uses two such symbols: a circle with a horizontal line underneath [ ⋅ ] and a long, strait mark similar to a comma (see figures 3 and 4 respectively). The first mark indicates the Swedish word och [and] while the comma-like mark stands for eller [or]. In these instances, the symbols are treated like abbreviations and resolved with the missing letters in italics. The first symbol, then, would be written “och” with the “ch” in italics. This is because many readers may not be familiar with the variety of Nensén’s abbreviations.

What these editorial decisions indicate is that context is important when determining if a mark should be treated as punctuation or an abbreviation to better facilitate readability. The same is true for two other examples which reveal how difficult the editorial decisions are to make: Nensén uses the connected letters “sk” to stand for the monetary value skillingar as well as another Old Swedish symbol to indicate the weight measurement lispund (see figure 5) that is no longer used today. In the first case, the group had debated whether or not to treat the grapheme as a symbol or as letters. Eckeryd (Personal communication, April 16, 2015) argued to treat them as letters. In the second case, the group was split between writing out the word in italics or by drawing a symbol in another programme to use in its place. The argument for using the symbol was that the group’s main aim was to keep the edition as diplomatic as possible. On the other hand Karlgren (Personal communication, April 16, 2015) reasons that, like the “sk”, it is too difficult for modern readers because the symbols are no longer used, and
this would distort the reading. Moreover, if the facsimile was set beside the transcription, this would allow the user to have the exact surrogate against a more readable version, and writing the symbol out would not harm the edition. Additionally, she states that other groups have attempted to insert a drawn version but with little success.

Figure 5 “Lispund” symbol, folio 19

Abbreviations were a significant aspect concerning the records as Nensén used them constantly (and not always consistently) to fit as much text onto a page as possible. There were both signs indicating abbreviations and the letters of the abbreviated word. Nensén also used underlines as in the case of letters “Bdr” to stand for bönder [farmers] (see figure 6). Editorial consensus was to resolve all abbreviations with the missing letters in italics to facilitate readability. The main issue, however, was between known and unknown abbreviations. Unknown being words where they were unsure about the word’s spelling or verb tense. Since at the time of the records the Swedish language was still influenced by Danish and Norwegian, Old Swedish may employ an “e” at the end of words where today it would use an “a” (R. Eckeryd, personal communication, April 16, 2015). Therefore, modernized spellings could be used in the case of abbreviations where the interpretation was questionable such as if due to unknown plural/singular tense or the definite/indefinite form of the word. But the transcribers followed the original spelling as much as possible.

Figure 6 “Bönder” abbreviation, folio 19           Figure 7 Three styles of “S”, folio 19: “sällan” and “Sidensjö”

Aspects involving letter forms included the script; typographic ligatures where two or more graphemes are joined together; and letter variations. As already mentioned the common handwriting of the time was Latin script, but Nensén had his own forms too. In the records “sk” and “st” are commonly connected, and the editors have identified five different styles of the letter “s” (see figure 7). The general consensus about the latter was that these were not to be distinguished in the transcription mainly due to tool functionality, as Eckeryd (Personal communication, April 16, 2015) maintains, “I don’t see how you could do that”. Karlgren (Personal communication, April 16, 2015) gives a good solution when she says, “I think if someone is interested in that they could search for the ‘s’ in the transcription and then look at the image, because we are going to have them side-by-side”. Despite the desire for as diplomatic a version viable, the editors did not have the means to represent the script or letter forms in Microsoft Word. Their argument was that the facsimile provided this if the user was interested. Their main goal was readability which was compromised by the mass of symbols, signs, and abbreviations.
The remaining features explored in the interview dealt with surface features and other properties of the records. It was agreed that all of these characteristics be kept as expressed to meet the aim of a diplomatic transcription; except in the case of headers. Headers were a more complicated element due to the layout in which Nensén formatted his folios. In some cases, the text runs vertically down the page. A discussion about this in the group interview revealed that although it was believed it was possible, albeit more difficult, to represent this feature in Microsoft Word it was not desired “because we want it in context” (S. Karlsgren, personal communication, April 16, 2015). Text that is placed outside of the main body may be more difficult for readers. The group has solved this issue by identifying what they move by using single tilting quotation marks to identify how the original text was presented.

Identifying other surface feature changes is also the case with resolving abbreviations. As the editors type out a line of text that was previously abbreviated (or with the addition of marginalia) it may extend the line longer than the Word document allows. Or in the case where the term “dito” (i.e. as stated in the line above) occurs, the editors have opted to insert square brackets with an equal sign and then the intended meaning. As line breaks are kept, this means that the transcription may run onto two or more lines. The group resolves this by formatting the lines closer together to indicate that they originally took the place of one. The transcribers also indicate if the insertion was due to marginalia so these two types of additions can be distinguished.

Another unique property to records is illustrations. The records held one (known) hand-drawn map. This presented a unique instance for transcription as the map included writing and critically interesting content. The editors had not come to a decision, but felt that this should be included as a facsimile image to keep it in context with the collection as a whole. Transcription of the map was more difficult, of course, but they thought that adding comments might be helpful to set places and names written on the folio to identify them in context.

The last property was document genesis, and this considered aspects such as document development and revisions; the type of paper and ink used; and other physical features that set the records in place and time. Here, the editors were clear that this was important, but not part of their task as transcribers. This job was delegated to those in the group deemed more fitting (i.e. contextualizing researchers) who were involved in the project and would write an introduction to the edition with this information (section 6.1.1).

6.1.5 Key themes

The themes identified in the analysis were context, diplomatic, readability, and simplicity. These will be explored more fully in relation to the tools, but first the thesis will look at these themes in relation to the goals, material, technical considerations, and transcription practices of the Nensén group.

A key theme that kept arising throughout the interview was context. Much of the discussion around the source documents revealed that context played a major role in transcription decisions regarding abbreviations, handwriting, and format. Context was
vital for first acquiring a feel of how Nensén wrote so that the editors could then transcribe the Sami material as neither knew the language. It was also important for diplomatic decisions such as preserving the excerpt material in the order it was recorded by Nensén or knowing that a section of text was missing. Editorially, this also played a role in actual transcription practice since the editor’s familiarity with the documents grew as they transcribed. It helped to resolve difficult or unclear words and abbreviations that could only be understood in context to the surrounding text. It was also true for solving other elements that were unclear through the biographical nature of the topics and in the case with deciding what constituted a heading.

Another aspect regarding context was how the transcribers used it to markup the text. Although they were not using traditional markup models, they were making critical decisions about how the text was to be understood by the reader. Using the comments function in Microsoft Word, they could expand, explain, or reference another instance in the text. These comments (later intended to guide XML-based encoding) were in themselves part of the critical edition, and sign-posted words or sentences that the editors thought important or of added value to the text as a whole. This practice indicated that the editors had an invested interest in the outcome of the edition as well as revealing the need for their expertise to assist both the later encoder and the reader. Context being important to transcription seems to indicate that a tool should allow for added material the editor deems important e.g. through the feasibility to include side notes, in-text references, or links to external addendum material.

Diplomatic was another theme recurrent throughout the entire group interview. As Karlgren (Personal communication, April 16, 2015) states plainly: “that’s the principle of it all: Diplomatic”. These concepts required editorial decisions to be set against what the tool allowed them to do and how it would affect the text. The transcription elements that were problematic for Microsoft Word included the map, the Latin script, symbols unique to Nensén, and various letter forms. Those elements that were able to be kept were spellings, capitalizations, certain punctuation, and surface features. In other words, graphemes were not so much of a problem for Microsoft Word; it was the graphic nature, or glyph, that could not be replicated.

Those properties, such as abbreviations and symbols, which were not kept and were not problematic, were done so because of readability. The tension between a readable text and a diplomatic edition were two themes that ran in conjunction with each other. Although they were not necessarily opposed (as in the instance where text formatting could allow vertical text or a symbol to be created), it was an issue of debate among the group members because of how each saw the final edition. It seemed from the transcribers’ views, that their main concern was that the user could access the text. They were clearly aware of their expertise as historians and their skills in palaeography but realized the average reader would not be helped by the addition of Old Swedish elements or various symbols. Here, though, the gap was narrowed by having the facsimile alongside the transcription and was one reason why such a presentation of the text was so vital to the edition. Karlgren (Personal communication, April 16, 2015) illustrates this when discussing the decision not to draw the weight measurement symbol. She says:
We could mark it at the beginning. We know what it is, it is a symbol used in his
time, but a lot of people going to read the sources and our description of it will get
stuck, and we think it is easier to read it in the text as a word than like this.

This gap is narrowed by presenting the reader with both the image and the text in an
accessible format. A diplomatic transcription is not entirely sacrificed when placed in
context with the surrogate facsimile.

Comments were another solution to resolving the tension between readability and a
diplomatic text. This function, as discussed in relation to context, allowed them to
compliment the failings of the tool functionality by clarifying and bringing the element
to the forefront. It seems to suggest that although a tool may hamper the exact
duplication of a document, it can be remedied by the transcribers who would be adding
this supplementary material any way.

Another key theme was simplicity. This was evident in that the Nensén group were
transcribing the source documents and nothing more apart from adding commentary
where necessary. With a project time of two years, this would need to be completed
efficiently as well. Therefore, Microsoft Word was chosen for the express purpose that
it was available, the two transcribers were familiar with it, and the project could begin
as quickly as possible (discussed in section 6.1.3).

As stated in the goals, it became clear that the transcribers were concerned only with the
transcription; that technical aspects were not their responsibility nor were the editors
particularly experts in software programmes. Such issues as cost, and tool integration
also influenced their choice to use Microsoft Word. Likewise, if they were to choose a
tool with specific transcription functions, one requirement might be that the tool was
quick to learn and easy to use. And specifically, functions such as XML encoding might
not be required of a tool by the Nensén group’s workflow.

However, Microsoft Word was not the most suitable choice for the Nensén editors.
They were unsure of how they would transfer the transcriptions to the online database or
if it would require more work to reformat and fix errors. Since integration was not an
issue for them to worry about they did not have a solution. Furthermore, they felt that
Microsoft Word was problematic in its enactment of transcription including formatting
as can be seen from table 1. Microsoft Word allowed them to perform only the simplest
transcription, but since no alternative was available, this was accepted.

The next two sections will address the second set of research questions: What are the
transcription tools available? What characteristics or features do they have that can be
used to review them? How should tool selection be approached?

### 6.2 Identification of tools

The tools chosen as being the most applicable for the Nensén group were
*FromThePage*, *Scripto*, and *Wikisource*. *T-Pen* was also of interest as it had been
previously named.
A transcription tool, as this thesis understands it, is an instrument that can enable the transcription of a handwritten text from a facsimile image into a machine-readable format by manually typing in the text the editor desires. In this instance, Microsoft Word qualifies as a transcription tool as it allows the Nensén group to do just that. Some of the tools in the list, however, only perform specific tasks or do not support certain types of documents. For instance, Xalan is a tool for the conversion of a text between markup formats, while ABBYY FineReader and ProofReadPage are OCR applications for printed documents and do not work for handwriting. The software CollateX allows the visualization of manuscript genesis, and PyBossa is a collaborative platform that enables micro-tasking such as annotation or image recognition. All of these aspects can be considered relevant to transcription and can be incorporated into other stages of a transcription project’s workflow, but not all are relevant or necessary for the Nensén editors and their specific documents.

Initially, 44 tools were identified through the chosen websites. Categories derived from the literature were used as the starting point for appropriate tool identification (see section 5.5). They consisted mostly of description elements and technical considerations; the latter being the preliminary factors in tool choice (Brumfield 2013). These included: the tool’s developer, if the tool is hosted; the license; the platform; if there is a cost involved; the text-type used by the tool; the document type supported by the tool; and content management system (CMS) integration where applicable.

These aspects were seen as important because digital tools are dependent on their technical infrastructure. Likewise, they are bound by their accessibility through cost and other considerations such as whether the license permits reuse, distribution, or even commercial derivatives from the software. For projects that may build upon or develop the software further, legal and copyright issues need to be considered.

Another aspect of tool functionality was whether or not the tool supported preservation and would be sustained as a digital tool. Many digitization projects hope to make their material available not only for users but as representative copies with long-term use in mind. Since digital projects are founded on the technology supporting them, this software needs to be supported in turn by developers so that those producing content need not be forced to repeatedly transcribe the material. Technical infrastructure needs to be sustainable and maintained if it is to avoid being made obsolete. Tools that are open source or widely used are more likely to be viable options for long-term projects as many more participants are involved. Also, tools that are interoperable or based on standards are more easily upgraded and therefore more likely to be valuable in the future.

This initial review revealed that transcription tools generally fall into one of two categories: audio/visual transcription or text/image transcription. This greatly reduced the number of tools to be studied further. Since the Nensén project specifically required a tool for text and/or image (facsimile) transcription, those tools handling audio or video transcription were directly removed from the list leaving 28 in total.

Other aspects were identified via literature on transcription practices and added in the form of three category fields: markup type (where applicable); other (properties specific to the tool including editing practices apart from transcription); and interface features. These were derived from a purpose-driven perspective about the tools and sought to aid
in functionality comparison. Many of the tools listed possessed specific functions such as tagging, annotation, side-by-side views, visualization, or zoom which could be added incentives or bonus features for projects wishing to build their digital collections.

The Nensén editors were only concerned with initial transcription, but they saw the benefits of having creative and unique functions enabling users to interact with the texts or for others, as historical and linguistic experts, to add supplementary material. Therefore, tools that allow for various interface options should be considered in case future developments, though they were not the deciding factor in selection.

Once the tools were found, they were then compiled into a table for better comparison (see Appendix II). Of these, all are considered applicable for projects in Alvin, as they can be seen to support the critical transmission of documents whether as steps in the process such as annotation or document conversion, but not all are applicable for transcription as the Nensén editors require it.

6.3 Selection of tools to be reviewed

Brumfield (2013) outlines a process for tool selection that considers the source material; the purpose of the transcript; financial and technical aspects; and the organizational fit. This final aspect is one which is of particular interest to the thesis. Since Alvin will ultimately be responsible for choosing a tool that can facilitate transcription to different groups, it is in their best interest that it not only facilitate scholarly editing practices, like the model provided by the group working with Nensén’s records, but that it also is simple and easy to use by those with little technical expertise.

Choosing a tool for a transcription project, according to Brumfield (2013), is made more complicated by the variety and variability of the tools actually available. Most especially technical considerations such as content management system (CMS) integration which may end up determining the outcome despite preference for a specific interface or ease of use. Any tool will require some customization and skilled expertise to be implemented for a project’s specific needs. Consequently, these facets are outside the scope of this thesis. The aim of the group interview was to identify the editor’s transcription model and explore how this was limited. As the results of the interview showed, separate roles were held by the editors and the university libraries (i.e. Umeå and Uppsala). The editors’ decisions stretched as far as copying out the text albeit with annotation and identification of important bibliographic keywords. Presentation, encoding, and user-functions remain the responsibility of the libraries and tool developers.

Since there was such a distinction between the roles of the editors working directly with transcription and the other project collaborators, it was considered valuable to review tools that had a range of specific editorial functions for the transcribers and not so much technical aspects which could be better utilized by those with the appropriate expertise. Nevertheless, after reviewing the results of the group interview, it was decided that tool selection for the Nensén group should emphasize certain generalized aspects that assisted in narrowing the selected tools. These fell under two main categories, the first being:
Simplicity: The tool should:

- Be free and open source (or similar licence), thereby supporting the organizational costs as well as internationally supported standards, collaboration, visibility, availability, and dissemination.
- Be easily integrated through a CMS: the data can be imported and exported into a new system.
- Support plain-text and facsimile images. Support for other languages or character sets.
- Be quick to learn and require no advanced programming or encoding skills. Markup is optional depending on the level of knowledge required.
- Be generally easy to use for those employing it. Make transcribing easy and/or have useful integrated interface features allowing the transcription model to be followed.
- Have any additional extra functions which may be helpful such as potential plugins for added functionality, collaborative multi-editor support, etc.

Features that supported the goals of the project, such as interface functions, text formatting, and output presentation and access, were considered and based on the properties found in the literature review of the tools (Appendix II).

In order to explore the tools in more detail, a second category of features was developed from the transcription practice itself, and thus became the basis for the review of the chosen tools. These focused on how the tool facilitated transcription to the editor’s standards. That is, it would be:

Diplomatic: The tool should facilitate the type of diplomatic edition the editors can achieve with Microsoft Word or provide a better, easier solution that considers:

- Orthography: the tool can provide desired punctuation and support different languages.
- Writing features
- Markup: if it assists with diplomatic features that are hampered by tool functionality such as letter forms or writing features and how these are expressed through the markup.

The results of the tool selection gave four tools that could be compared based on how they fit the first category of simplicity, and then how they performed transcription to diplomatic standards. The chosen tools were: *FromThePage*, *Scripto*, *Omeka*, and *Wikisource*. These also presented slightly unique features. However, it was clear that *Omeka* could be integrated as the content management system (CMS) for *Scripto*. Therefore, both these reviews could be combined. It was also clear that *WikiSource*, *Scripto*, and *FromThePage* are all based on wiki markup. This presented a problem in that *T-Pen* had been named as a possible tool because more advanced transcriptions using the TEI guidelines were desired (Per Cullhed, personal communication, January 22, 2015). In this case, it was considered valuable to review *T-Pen* for transcription purposes as there was a distinction between the library’s role and the transcribers’ role. Whether the decision to use it is made or not, reviewing it could help identify features that are valuable to consider in regards to the other tools.
6.4 Tool Review

This section will present the tool review in answer to the third set of research questions: How do the tools facilitate transcription? How do they support or hinder the transcription practices undertaken by the group? Do tool features meet the needs of the editors?

The tools chosen for review are those that meet the general criteria of simplicity in how they might be implemented, and subsequently, are explored in how they realize a diplomatic edition to see if the model fits that of the editors working with the Nensén records. A quick synopsis of each potential tool will provide an introduction to its suitability.

6.4.1 FromThePage

*FromThePage* (http://beta.fromthepage.com/; http://www.fromthepage.org/) is a free and open source transcription tool and was designed out of personal interest by software developer Ben Brumfield for the collaborative transcription of his great-great grandmother Julia Brumfield’s 1922 diary. Since its development in 2007, *FromThePage* has been used by a range of other cultural institutions; one example being The San Diego Museum of Natural History which is working on a project to transcribe the field notes of herpetologist Lawrence Klauber.

*FromThePage* is built to use the syntax-structured wikilinks markup. This allows subject identification for building an index or database of keywords and facilitates full-text searching (Brumfield 2014a). Brumfield (2014a) also indicates that wikilinks (see section 6.7) is a valuable tool because it enables amateurs to export a text document as a TEI P5-compliant XML file. It also claims to be unique among tools in that it can link in-text words to subject categories as well as automatically suggesting links through markup mining.

Being created as a personal project is its strength for users of different skill sets since *FromThePage* is well-documented through Brumfield’s blog *Collaborative Manuscript Transcription* (http://manuscripttranscription.blogspot.se/), and he remains an active participant in discussions and conferences. *FromThePage* originally utilized, the now inactive, *Internet Archive* as its content management system (CMS) but is currently being developed for integration in *Omeka*. It has also expressed an interest to collaborate with *Fedora Commons* (http://fedorarepository.org/) (Brumfield 2014b), so the transcription tool remains an active and developing possibility.

This tool is one option for the Nensén group as wikilinks may support their division between transcriber and encoder. It could assist in creating a type of markup to distinguish important bibliographic information that would also be used for indexing the texts to whatever capacity the editors deemed necessary. It would be directed by the historical expertise of the editors themselves, while encoding expertise would still remain the role of the library.
6.4.2 Scripto

Scripto (http://scripto.org/) is an open source transcription tool developed by the Roy Rosenzweig Center for History and New Media at George Mason University. It is designed to fit any pre-existing CMS incorporating documents, images and multimedia, but provides guided instructions for using Omeka (also developed by the Center for History and New Media), WordPress, or Drupal as the CMS publishing platform.

Scripto is specifically designed for libraries and museums working with collaborative digitization projects in the humanities on any scale, and can even be used by small groups or individuals. In this capacity, it authorizes two types of accounts to be created: editors and transcribers. This is considered beneficial for crowdsourcing as large projects can be completed by transcribers and the transcriptions can be quality checked by the editors. The editors have the same functions as transcribers and thus the tool can be adopted for private projects.

As a MediaWiki-based tool, Scripto can be used as is, or adapted to specific projects as in the case of DIY History (http://diyhistory.lib.uiowa.edu/) from the Library at the University of Illinois Urbana-Champaign among a great many others. An example of a commonly utilized plugin includes the ProofRead Page extension\(^3\). However, extension of the transcription tool would be dependent on the skills of those implementing it and not the editors themselves. These choices would be driven by additional, long-term goals for the Nensén project.

Omeka (http://omeka.org/) was identified as a possible digital editing tool, but since it also assists Scripto as a CMS and is widely employed as the chosen publishing platform for other cultural heritage projects, it is worth looking at in conjunction with Scripto as more applicable than WordPress or Drupal.

What makes Omeka a valuable publishing platform is that it can be extended through plugins developed especially for the open source CMS. This enables it to be adapted for digital repositories and collection management. It claims an easy-to-use interface and setup with documentation and support through discussion forums. A wide variety of libraries, museums, and archives are adapting it to their digital cultural heritage projects and collections. Two features useful for these types of material are Dublin Core standards for metadata and Unicode UTF-8; both supported by Omeka.

6.4.3 Wikisource

Wikisource (https://en.wikisource.org/wiki/Main_Page), developed by the Wikimedia Foundation, like FromThePage and Scripto, is an open source collaborative transcription tool designed to build library collections in many languages and translations. An interesting circumstance about Wikisource is that it was used successfully by Les Archives départementales des Alpes-Maritimes to transcribe 17th-

\(^3\) ProofRead Page (http://www.mediawiki.org/wiki/Extension:Proofread_Page) renders OCR text of a facsimile or divides the facsimile into its organizational and textual constituents. Such a feature could be useful for formatting handwritten manuscripts as OCR is still limited to print materials. This may not be practical for the Nensén group.
and 18th-century records of episcopal visits to communes in the diocese. This project is elucidated more effectively by Brumfield (2012a). However strikingly similar this project is to the goals of the group transcribing J. A. Nensén’s records, it is also a useful tool regardless of similar applications.

*Wikisource* uses the *DjVu* file format which stores scanned images containing combinations of images and text. Like *Scripto*, it is built on plugins for editing, most notably the MediaWiki *ProofRead Page* which allows proofreading of the OCR’d text to take place side-by-side the facsimile for comparison.

*Wikisource* aims to provide public domain material in digital editions and supports over 60 languages. As a free, collaborative effort, documentation exists in the form of discussion and a community portal to assist in implementation and use. However, because it is built from public domain and free-of-copyright material, it has a strict policy of what can be included in the archive. Another issue is that content is created by many users. Despite this, it is widely adopted and used as an editing tool which has greatly increased the accessibility and visibility of material institutionally housed worldwide.

### 6.4.4 T-Pen

*T-Pen* (http://t-pen.org/TPEN/) has been previously identified as one of the potential transcription tools, but is not currently being used by *Alvin* due to problematic integration issues and because the Nensén group does not have the skills needed for implementation. However, it may be worth reviewing to gain an idea of why it was considered at all and how its functionality might guide the choice of another tool. This thesis has chosen to review *T-Pen* because it might serve as a future guide if Uppsala decides *T-Pen* is the appropriate solution or if other groups have the means of implementing it in their projects.

*T-Pen* follows the defined pattern of a free and open source collaborative tool. It was developed by the Center for Digital Theology at Saint Louis University specifically for medieval manuscripts, and it is employed world-wide by many institutions. The tool is designed for collaborative transcription in two capacities: document-based transcription (where the editor can work on a single document) and project-based transcription, the latter permitting a project management component for personalization. This enables groups or individuals to develop projects in the capacity they desire. Technical and editorial support and documentation are provided through the user-account and via the *T-Pen* blog (http://digital-editor.blogspot.se/).

*T-Pen’s* aim is to facilitate the creation and manipulation of transcriptions with multiple views. One function claims to allow exports of transcriptions as a PDF in plaintext XML for processing. Similarly, existing text, image, and data storing standards are supported. As a palaeographic tool geared towards medieval manuscripts, *T-Pen* includes character sets to support these types of documents in various medieval languages as well as providing Dublin Core metadata or the ability for Dublin Core or TEI extensions for specialized projects.
6.5 “Keep it simple”

Simplicity was one of the themes discovered in the interview and one which greatly effects tool usability and application.

Of the four transcription tools reviewed, all fit the goals of simplicity as they adhere to the characteristics of free and open source. One additional similarity between each of them is that they are classified as collaborative tools. But this need not be a problem, rather a further indication of the tools’ suitability.

Brumfield (2014b) makes a point that collaborative tools do not necessitate collaborative projects and can function as well for private projects. One benefit of collaborative tools is that they are usually designed for amateur users, as was the Transcription Desk for University College London’s Transcribe Bentham project (Causer, Tonra, & Wallace 2012), and so they are typically easy to use and quick to learn. Markup is generally wiki-based and subject-centric so projects can be instigated quickly and efficiently by providing online manuals through blogs or wikis. Collaborative tools also tend to be open source tools. This is beneficial because they are widely supported and therefore the tools are more likely to be updated and the documents more likely to be viewed and shared (Buchanan 2014).

Nevertheless, simplicity also extends to integration and the technology needed to implement them in projects. As in the case with Scripto, Buchanan (2014) identifies a few problems. The main complication being that the most feasible publishing platforms are Omeka, WordPress, and Drupal; a project with a different CMS might have implementation difficulties. At the same time, though, T-Pen has no integration for these more common platforms and requires very specialized expertise to handle importation and exportation of data. Buchanan (2014) also notes that Scripto requires some level of technological skill to implement it as it is heavily based on extensions. Certain plugin functions, such as Microsoft’s Zoom.it, are subject to updates rendering them unavailable during periods. Extensions are an issue with both FromThePage and especially Wikisource. Specialized projects, however, are subject to unique considerations as seen in the interview with the Nensén group. Any internal or crucial project goals will have to be considered when deciding what additional functionality is necessary. Nevertheless, the MediaWiki (http://www.mediawiki.org/wiki/MediaWiki) plugin enables full-text searching to be rendered in transcriptions and is therefore of considerable value in tools such as Scripto or Wikisource. While in FromThePage, wikilinks ensures that this goal is met.

Another useful result of the tools’ compatibility with plugins (e.g. ProofRead Page) is a very helpful solution for addressing document formatting. Scripto especially, indicates that the transcription tool does not format the document. Any diplomatic edition, therefore, is hampered by this if surface feature accuracy is desired in the final transcription. Wiki markup (http://www.mediawiki.org/wiki/Help:Formatting), explored further in section 6.7, is one possibility that can enable transcribers to approximate the original document formatting. As an extension identified by Scripto, MediaWiki is also a means of providing plain and simple markup to encode the document since it does not

---

4 For an example of a wiki page, see the Transcribe Bentham’s wiki page http://www.transcribe-bentham.da.ulcc.ac.uk/td/Using_the_transcription_interface
immediately support this. Brumfield (2010) discusses the ProofRead Page software’s ability to organize and display the document in the appropriate context because XML is poorly suited for overlapping textual structure.

Considering actual transcription practice through the tool’s interface is another important aspect of simplicity as it renders the tool usable or not. All the tools feature a transcription box beside the facsimile with a zoom function for detailed viewing while transcription occurs in free-form text. Encoding or markup is not necessary for the editor unless considered of importance. Even so, since the tools are collaborative, they provide a means of allowing others to edit the text. This enables each transcriber to perform the task to which he or she is best suited.

One interface characteristic that should be mentioned as unique to T-Pen and not the others is that transcription is performed line-by-line (see figure 8). T-Pen is able to automatically parse 80% of folio columns, while manual column parsing allows for the remaining folios to be transcribed appropriately, and therefore the tool is able to link the transcription to the same line in the facsimile image. FromThePage, Scripto, and Wikisource (see figures 9, 10, and 11 respectively) are page-centric and would require a plugin to facilitate this type of linking.

Figure 8 T-Pen, the Center for Digital Theology at Saint Louis University

---

5 It should be mentioned that all screenshots are my own, and they are only meant to provide a visual aid for the various interfaces. All transcription examples shown are the work of the projects to which they belong.
Friday, January 3, 1919

[Benjamin Franklin Brumfield, Sr. [Dead] worked on [tenant houses; Ozone house]]
all day. [Sally Joseph Carr Brumfield; Jesse]
[powing; wood] canoe. I got [dinner]. Boiled [hops; feet].
This evening [Benjamin Franklin Brumfield, Jr. [Frankie]]
[failing; feet] out of the window. It hurt him
very bad not to break any bones.
There is a [dance] at [Dick Wood; Dick
mangle] to night, Ben J. Jones
and [Virginia Harvey; Virginia] and [Irvin Harvey]
has all gone down there.
8 o'clock.
These other tool interfaces are similar in structure by providing the editor with a window in which to transcribe. Scripto is unique in that the developers have specifically designed the interface to be displayed vertically (Albers 2011). Each tool also allows for a preview of the work completed and a saved history of each version so all changes to the transcription can be accessed. This would allow the editors to review each other’s work or return to a different version at any time. In Scripto, wiki markup guidelines are provided by way of a link to the wiki page, while Wikisource offers this as a number of buttons to render formatting in a more automated way (see figure 12). FromThePage does not seem to offer help for formatting; reference to the MediaWiki guidelines remains the editor’s choice and is dependent on his or her expertise.

The automated functionality of the tool bar is one feature that could be why T-Pen was singled out (see figure 13). By clicking one of the buttons, an editor can add a character, XML tag, annotation, or abbreviation. He or she can also check a Latin dictionary alongside several other contextual functions like viewing revisions or pages.

FromThePage and Scripto do not have a toolbar and so certain in-text additions must be keyed manually, or else the editor must forgo these altogether. Wikisource does have such a toolbar (figure 12); it being one advantageous aspect to the tool’s otherwise convoluted and unclear interface design. The Wikisource toolbar functions something
akin to the Microsoft Word toolbar having formatting buttons including the ability to automatically insert headings, paragraphs, images, links or other references.

Figure 13 T-Pen Toolbar

What is quickly realized is that interface design affects both the simplicity and the diplomatic level to which a text can be transcribed. The available functions aid the editor and simplify the task at the same time as allowing these diplomatic decisions to be made at all.

6.6 “That’s the principle of it all: Diplomatic”

Diplomatic was another theme discovered in the group interview. It is closely tied to readability and so these two will be treated together and reviewing the tools in light of how they enact these themes is important to their functionality.

The properties required for a diplomatic edition are the decision of the editors. No digital transcription of a text can reproduce the manuscript in every detail, but to some extent the tool must facilitate the editorial transcription model considered of value for the desired edition as discussed in the problem description and in editorial practice (see sections 2.3 and 3.4 respectively).

The transcribers working on the Nensén records wish to keep the edition as diplomatic as possible; much of which is hindered by Microsoft Word’s capabilities. Additionally, the reviewed tools also pose restrictions on how far editorial decisions can be enacted on the text, and therefore, the choice of tool will limit the final text to some extent.

To maintain diplomatic orthography and writing features requires that the tool, to some degree, support punctuation, capitalization, symbols, sub/superscripts, and especially character sets in the languages the manuscripts are written in. In the case of Nensén’s records, this is important as much of the research value comes from J. A. Nensén’s diligence in documenting vocabularies in various languages as well as terms that are no longer in use. Thankfully, this is not too problematic as all the tools support this to some extent (e.g. Omeka supports Unicode UTF-8 on imported records). For the most part, the editor can transcribe the text as desired using the keyboard, but for inserting specific or unique glyphs, the toolbars are better suited. On its website, Scripto (http://scripto.org/documentation/appendices/) specifically states that only symbols replicated by the keyboard can be entered, otherwise transcription guidelines recommend a description of the character in square brackets.

Since T-Pen was created for medieval manuscripts, it allows the addition of medieval letters, e.g. the Old English thorn [þ], eth [ð], and wynn [Ƿ]. T-Pen also allows the editor to add character extensions if needed. Wikisource, on the other hand, uses a “special
characters” set (see figure 14) including the extended Latin alphabet, various phonetic alphabets, and other diacritics and symbols.

In regards to letter forms, T-Pen can display digraphs (e.g. æ), while Wikisource can render, for example, the long “s”. It is unclear how FromThePage and Scripto can render these types of instances. Since they are based on extensions, a plugin may rectify this, or else notation by the editor must indicate a feature that is not documented as expressed in the source document.

In regards to surface features, the tools are more difficult to assess. That is, FromThePage and Scripto do not have toolbars to insert line/page breaks, or format headers or paragraphs. These two tools work through free-form transcription. This is a kind of “what you see is what you get” (WYSIWYG) interface that roughly formats texts as they are transcribed. As can be seen from figure 15, transcription is made simpler so the editor does not need to focus on additional tasks. This may actually aid transcription by removing extra steps interrupting the reading flow or contextual nature of transcribing. However, if columns or lists are required, wiki markup is necessary to render these characteristics. Scripto specifically gives the advice to use two hard returns to indicate a line between paragraphs. This would indicate a type of internal and invisible formatting such as a WYSIWYG that underlies Scripto.

Other properties to be reviewed are abbreviations, annotation, and genesis. As was discovered in the group interview, abbreviations for the Nensén records were expanded due to readability and not directed by tool functionality as some of the other elements. This means that such a feature is not necessary in the tool itself. If the editors were interested in indicating where an abbreviation occurred, they could either do this in the same way as they approached this issue in Microsoft Word (i.e. using square brackets), or else use markup. This would be possible only in T-Pen, though, where XML is employed for document encoding. Wiki markup is for formatting text and would not be suitable. However, this is not the current intention for the Nensén transcriptions.
This thesis has decided to include annotations and links as properties in conjunction with accidental features for three reasons. First, since all the tools facilitated annotations (and/or comments through wiki markup) it was considered useful to mention. Second, as comments were important to the transcribers in the Nensén project, it would be helpful for them to have some type of instrument, however simple, that could meet this need. And third, comments and annotations also support the detailing of document genesis. None of the tools have any specific function for chronicling the manuscript’s creation. Though this is mainly done through encoding, annotations could be one option for allowing some extra detail to be marked by those with limited knowledge of XML. And whether or not this is entirely necessary is up to the editors to decide. One recommendation is that the transcribers have a certain level of expertise regarding the records and this is aided by their detailed analysis of context during transcription. If they saw fit to record more information about the document such as a missing page or a possible addition of text in a place previously left blank by Nensén, this information would be of value for later scholars or others involved with historical research on the records.

6.7 Markup

One theme found in the analysis of the group interview was context. It has specific implications regarding the Nensén records as it can be viewed as a foundation to markup. That is, though the group will not traditionally add markup to the texts, they will be guiding future markup through side notes and contextual references to aid indexing and textual structure to the manuscripts. Since markup brings structure to the

---

As an aside, it might be either helpful or superfluous to add that annotation is a separate scholarly primitive according to Unsworth (2000). The tools designed for this purpose make up a group of their own and should be reviewed separately.
text and supports indexing, linking, and full-text searching, it can be seen as a task dependent on the contextual understanding of the documents.

This section will not attempt to discuss the differences between wiki markup and XML (or the TEI guidelines), rather it will attempt to expound on the benefits and limitations of each in a general way in order that it can assist guiding selection of the most appropriate tool.

As can be seen from the review, the first three tools use wiki markup to some extent. However, *T-Pen* was considered the tool of choice and it is based on XML/TEI for document encoding. The Nensén group has also stated that markup is desired and will be added at a later date by the scholars developing the digital archive. Markup and encoding through standardized formats, such as with the TEI, is a very useful and vital instrument in which to enable the long-term preservation and accessibility of digital documents (Burnard, O’Keeffe, & Unsworth 2007b). Specifically, the TEI is intended for humanities texts and so is designed with a better foundation in which to express these types of documents in addition to being adaptable (Hockey 1999, p. 8). Therefore, this thesis considers it essential to touch on this topic, if only superficially.

Through these tools, markup is used for encoding the document in two ways: First, for formatting, and second, for recording meta-information. Formatting the text has already been discussed in regards to interface design and the diplomatic nature of editions such as adding headings and organizing the textual structure of the edition (sections 6.5 and 6.6 respectively). Indeed Causer, Tonra, and Wallace (2012, p.123-124) discuss how Microsoft Word’s strengths are in its formatting abilities, while its pitfalls are that this formatting does not identify the function or meaning of the text as markup can do.

But this formatting is enacted by markup in both *Scripto* and *Wikisource*, that is, the tools that can use wiki markup. This is because wiki markup is designed to format the text instead of recording semantic substantives and textual components as the TEI is intended for. Wiki markup includes some basic HTML tags and symbol entities that allow a text to be presented onscreen. This, of course, enables the tools with limited functionality to display a text with more diplomatic features.

The other use for markup is the more traditional practice of encoding a text with XML for machine readability and documentary organization. The TEI (http://www.tei-c.org/Guidelines/) lay out a set of guidelines that define the “structural, renditional, and conceptual features of texts”. Using XML, the TEI can organize a text into its structural components and provide tags defining semantic properties and elements such as names, places, and dates. This allows the text to be parsed in order to search or extract indexes. *T-Pen* is a tool that provides this type of markup for the text.

But there are two main problems with markup. First, it is difficult; the complexity and skill required make for a strong argument against it (Lavagnino 2007). The Nensén group identified that they were not the ones to encode the texts as their task is plain transcription. They felt they did not have the appropriate skills, nor was it in line with their goals. But this can be problematic for the continuity of the project. Ginther (2012) relates that many projects have a three step process which sees these roles divided between separate parties. However, since context is vital to transcription, it would be
more helpful to provide some way for the transcribers to encode the text, in whatever capacity.

*T-Pen* has addressed this by creating a toolbar with XML elements incorporated into the character stream (Ginther 2012). This can be modified by linking an XML schema to the project by the project leader. For an example of XML tags in *T-Pen*, see figure 16 from the *Carolingian Cannon Law Project* (http://ccl.rch.uky.edu/). However, there is no distinction between metadata elements and those found in the encoding schema so editors must remove those elements not being used.

![Figure 16 T-Pen XML tags, example from the Carolingian Canon Law Project](image)

University College London’s *Transcribe Bentham* project has approached encoding from another perspective. Their *Transcription Desk* (http://www.transcribe-bentham.da.ulcc.ac.uk/td/Transcribe_Bentham) runs a MediWiki and therefore, the project was able to develop a plugin for encoding. The *Transcription Desk* uses a TEI toolbar (see figure 17) which allows editors to insert tag sets by clicking on a button. This basic tag set enables amateur editors to produce XML-valid transcriptions. One benefit of this collaborative and open source project is that *Transcribe Bentham* has made the TEI toolbar source code freely available for other transcription projects (Causer 2011). This could be a potential option for projects using wiki-based tools if basic TEI encoding is preferred to wiki markup.

![Figure 17 Transcription Desk Toolbar, University College London, Transcribe Bentham Project](image)

The second problem with markup, specifically XML, is that it is not always suited to documentary structure (Eggert 2009, p.68). Elements in the text overlap, while XML tags cannot. However, this may not be an issue with the Nensén records as they are typically well-structured. What might be problematic is if a section of text spans two pages. Brumfield (2010) considers this problem partially addressed by the *ProofRead Page* extension which uses hyperlinks and annotation when a line crosses a page break.

*FromThePage*, however, has approached the need for semantic indexing using an entirely different markup from the other tools. Brumfield (2014a) describes how using wikilinks can create indexes and allows the computer to call up pages and references. Wikilinks is considered simpler for users as it is subject-based and thus the syntax is more straightforward and in plain-text. It allows the editor to mark up bibliographic information by identifying them as they are shown in the text at the same time as connected to their canonical form so elements such as initials can be expanded and links
can be created between a name and its variations. This feature is useful not only as it enables the computer to automatically suggest markup, but can mine the text for subject categories. Wikilinks is also useful for exporting as an XML file which supports interchange of data. Wikilinks, though, may be more beneficial to smaller projects requiring minimum markup and formatting. But such a solution to the various uses and difficulties with traditional markup can be met through this method.

Though markup is perhaps not entirely within the scope of the Nensén transcriptions, by giving a small introduction to the tools’ applications of encoding possibilities, this may assist in clarifying the benefits of bringing this task into the realm of scholarly editing and transcription.

6.8 Tool transcription models

To conclude the discussion, table 2 has been created to illustrate the tools’ transcription models based on the review. It seeks to assemble the results of the review by offering a picture of how each tool expresses diplomatic features and enacted transcription to create a model for each tool:
<table>
<thead>
<tr>
<th>Theme</th>
<th>Property</th>
<th>FromThePage</th>
<th>Scripto/Omeka</th>
<th>Wikisource</th>
<th>T-Pen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orthography</td>
<td>spelling</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>diacritics</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Writing features</td>
<td>capitalizations</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>punctuation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>symbols</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>superscripts</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>?</td>
</tr>
<tr>
<td>Letter forms</td>
<td>text/script</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>ligatures</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>letter variations</td>
<td>No</td>
<td>No</td>
<td>Some</td>
<td>No</td>
</tr>
<tr>
<td>Surface features</td>
<td>line/page breaks</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>paragraphs</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>columns/lists</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>headers</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Abbreviations</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Other</td>
<td>annotations/links</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>genesis</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Markup</td>
<td>wiki markup, wikilinks</td>
<td></td>
<td>wiki markup</td>
<td>wiki markup</td>
<td>XML</td>
</tr>
</tbody>
</table>
7 Discussion

This thesis aimed to answer the questions pertaining to available transcription tools and how they facilitated transcription of cultural heritage material into the digital medium by looking at the practices of editors working with records by Lapland priest J. A. Nensén in a project to publish them through the Swedish library database Alvin (see sections 1.3, 1.4, and 2.3).

All four of the tools that were reviewed adhere to the free and open source model. They are all designed for collaborative transcription projects and widely used by cultural heritage institutions, libraries, museums, and individual groups worldwide. This made for an even ground on which to assess their merits. These tools were FromThePage, Scripto, Wikisource, and T-Pen (see section 7.1), and were explored based on the identified characteristics of simplicity (i.e. technical infrastructure) and how, and to what extent, diplomatic transcription was enacted (see sections 6.5 and 6.6 respectively).

The main finding confirmed the reasoning behind relinquishing T-Pen as the project tool: that technical and system integration determines the tool over usability or appropriateness for transcription. That is, although T-Pen provides many automated functions (making transcription and markup easier), as well as XML-based markup (desirable for standardized, machine-readable content), if importing and exporting data is problematic or not possible, Uppsala will have no choice but to use another tool. This, however, need not be too much of a setback as the other tools provide almost equal functionality albeit in other ways.

Currently, the choice to use Microsoft Word is problematic for the Nensén group as it is unsuitable due to poor integration. Exporting data is difficult with this tool: this being a main step towards viable online publication.

Further issues with Microsoft Word were that it limited certain desired transcription elements such as symbols that were not in opposition to readability. In essence, all decisions regarding readability would also be made in any of the other identified tools (as none of them supported more advanced features to address these issues) and so Microsoft Word was satisfactory in this sense.

Readability was the most desired aspect of transcription and supported the argument that an exhaustive transcription is not possible as previously discussed by Huitfeldt and Sperberg-McQueen (2008, p. 307) nor indeed desirable in a digital text. It furthers the understanding of transcription as an interpretation (Sperberg-McQueen 2009; Pierazzo 2011) that is both necessary in digital textual reproduction and acceptable. From this thesis, it can be argued that interpretation is, in fact, beneficial to digital texts that are bound by technical limitations such as problems reproducing symbols or letter forms as it helps resolve the disparity between readability and diplomatic that would be present regardless of the medium or digital tool chosen.

The main limitation with Microsoft Word was formatting, as that did not affect readability as dramatically, and therefore was considered important to preserve and maintained the diplomatic text in this respect. This was the case with all the tools and each tool enacted formatting differently. This directly supported the fallacy of the
second assumption discussed by Huitfeld and Sperberg-McQueen (2008, p. 308) by pointing out that in practice different markup systems have different notions of textual structure. This could be seen in how the Nensén group approached headings in Microsoft Word, but could potentially make a different decision if using another tool that supported more advanced options.

Of the reviewed tools, all of them present benefits and limitations for the Nensén group. Scripto can be seen as a viable transcription tool for several reasons. The main reason being its successful and established use in many other projects which provides it with a strong foundation as a useful and reliable tool that remains supported by developers and users alike. Another reason is that the content management system Omeka has been developed for similar purposes and provides a free and simple publishing platform for editors of cultural heritage materials. As well as supporting library and museum metadata standards, Omeka can be adapted for specific projects with an extensive list of available plugins designed especially for it. Moreover, Scripto makes transcription possible through a clean, user-friendly interface. Transcription is in plain-text with the choice to format the text with wiki markup left up to the editor. Wiki markup is also developed to be convenient and functional. Based on the goals of the Nensén project and the editors, Scripto is a possible choice enabling them to transcribe the documents adhering to the transcription model.

But since the goals of the Nensén group are the accessibility and usability of the documents beyond plain transcription, the tool needs to facilitate this. In regards to markup, XML or the TEI guidelines have been mentioned as the ideal. Here, the goal is preservation. Whoever intends to encode the text, be it library staff or editors, will need to ensure the markup used supports this long-term objective. Consideration here is important as wiki markup may not be entirely suitable for this objective.

Limitations with Scripto and FromThePage, apart from not supporting XML, are that they are not automated. Unlike T-Pen, which makes the transcriber’s task very efficient and easy, these tools may require more work on the part of the editor in manual transcription. The question then is: would Scripto or FromThePage require more work than is already being done by manually typing out the transcription in Microsoft Word? Two potential possibilities to promote continuity here exist. The first is using FromThePage’s approach with wikilinks markup for semantic encoding that can automatically generate indexes, full-text search, and computer-suggested markup to the extent the editors want. Wikilinks markup would assist with linking between references in the text, as well, and would facilitate data export in XML-valid documents. The second option could be to use the freely available TEI toolbar plugin developed by the Transcription Desk (Causer 2011). Collaborative projects benefit greatly from the input and development by other institutional projects, and this could allow the Nensén group to encode with simple TEI tags themselves.

Of course, no tool is without the need for some technical proficiency on the part of the users to implement it. What makes the need for simplicity in tool design complex is the dominance of integration requirements and skill of the system administrators. Set against the need for an easy, usable interface, project amalgamation complicates matters when disparities in roles and expertise exist. Collaboration, here, is key. Its benefit for the Nensén group is that these roles can be undertaken by the appropriate individuals (i.e. library staff at Umeå University or Uppsala University). However, communication
and a clear understanding of goals needs to exist to enable the administrators to pick the tool best suited for the project as well as being helpful for the transcribers’ personal vision for the Nensén records.

As to their ability to enact transcription to the Nensén group’s standards, all the tools, allow the editors to enact the majority of what they desire for a diplomatic transcription. Their limitations lie in formatting and the confinement to keyboard-based writing features, mainly, in configuration of columns and special symbol representation. The problems confronted by the editors in their current practice were also these. And so, using a tool such as Scripto is beneficial in that simple transcription is easily facilitated and that it allows for effective publishing; it could make the materials accessible in a user-friendly platform.

Considering the other aspects, both FromThePage and Wikisource are similar in how they address transcription. FromThePage, though, is hindered by the uncertainty in how much technical support it can provide editors, or in its stability as a preservation option due to its development for a single project. Such a tool may not have appropriate funding or interest for further implementation. Wikisource’s limitations, in a different capacity, were more centred on its usability and technical infrastructure. It has a complicated and convoluted interface which requires skill to navigate. Though similarly, in that it restrains the material acceptable for publication and allows general users to act as editors with insufficient quality-checks, Wikisource may not support long-term preservation either.

Challengingly, the tool best suited to the Nensén project’s actual transcription practices was T-Pen. It remains the most beneficial as it meets the transcription model laid out by the editors and facilitates more automated work and TEI support. However, since it requires solutions to integration problems that are beyond the Nensén editor’s skill set, it is not appropriate in regards to simplicity. As stated earlier in the discussion, tool usability and integration are more vital to selection than how the tool enacts the desired transcription model.

Since traditional transcription practice as the representation of a text has carried with it many assumptions about how a text should (and can) be represented, critical digital transmission of these texts have only been conveyers of these problems as pointed out by Robinson (2005). Additionally, they have created more issues regarding technical infrastructure that falls outside the roles of the scholars editing and transcribing these texts. Editors are well versed in presumptions about textuality and seek to address better ways of transcribing texts while representative efforts via new media have not. Ill-founded beliefs about technology’s infallibility or novel methods of performing traditional tasks on the part of developers only complicate matters and facilitate assumptions.

Libraries hold a privileged place in critical digitization efforts. They understand the need for cooperative efforts between editors and software developers to better enable the production and accessibility of digital materials. Many digital tools are built with little understanding about the needs of editors or transcription and they enact ideas about text that ultimately inhibit or limit how these texts are used. Libraries can work to bridge the gap between these two actors to better facility critical digital transmission.
8 Conclusion

This thesis used a semi-structured group interview to approach discussion about transcription as critical transmission that sought to better understand it in the digital medium. It then performed a tool review to place the two sets of findings alongside each other.

It can be said that the conclusions of this thesis are that:

1. Tool selection is dependent on the desired transcription model, the goals of the project, but most especially on the skills of the editors.

2. Digital transcription tools vary widely in their functionality and limitations. No tool can be deemed fully appropriate for all transcription projects. Technical simplicity ultimately determines the most usable tool.

3. The transcription model is very much interpretive in its own right, but also these editorial decisions on readability and diplomatic ideals are dependent on tool functionality.

4. Textual interpretation is beneficial to digital transcription because the limitations of digital tools may hamper diplomatic decisions or readability. Interpretation can lessen this gap.

5. Libraries are key actors in critical digitization projects because they can bridge the gap between editors and tool developers enabling a better understanding of digital transcription to improve tool functionality.

As discussed in Chapter 3, the tradition of this editorial practice carries assumptions about the text when it is transferred into a digital document. These were mainly two: first, transcription is exhaustive in representing the source document and, second that documents are structurally simple and therefore easily represented through markup thus making them accessible electronically. The results of the interview support the fallacy of these assumptions in that it highlights first, the interpretive nature of transcription; it necessitates choices on what to represent. And second, that representation (both structurally and contextually) of the documents was guided by tool capability.

The main dichotomy, as seen in practice via the group interview, was between making the transcription diplomatic and readable; the editors were required to judge for each instance which elements needed to be retained and which must be sacrificed in place of more desirable components based on goals. All of which further supports the need for tools to be developed and selected based on goals for digitization projects.

The two themes, readability and diplomatic, spanned both the editorial practice and the tools themselves. Context was also relevant to both aspects as so many editorial decisions influenced the relationship between the text and its digital reproduction.

Simplicity, however, was an important theme in that it was solely a consideration in tool functionality: the editorial practice was anything but simple. Nonetheless, simplicity
was a profound issue in tool use because the editors were bound by their technical abilities. No matter how well a tool is able to enact the transcription model, if the editors cannot use it, it is poorly designed. This aspect draws more light on the need for software developers to build tools that are usable from the editor’s perspective.

Such a conclusion is useful to libraries undertaking critical transmission as libraries stand at a junction between their collections and the users who derive value from them. On the one side, they act as collectors, repositories, and preservers of unique and significant cultural heritage materials that may be physically restricted due to location, condition, or simply the lack of a catalogue record. On the other, and perhaps more importantly, they are the distributors of this information since they seek to make their holdings accessible, available, and functional for scholarly purposes. Many of these holdings are made so through digitization initiatives by the library to which they belong or through collaborative projects with other experts.

Digitization itself is an interdisciplinary endeavour drawing from humanities studies, computing, information management, and library and information science thus making libraries ideal participants to be at the forefront of these initiatives. Libraries, such as in Uppsala University, stand to gain by a review of the tools. Editorial groups involved in institutional projects, like Alvin, could benefit from the library’s support in choosing the appropriate tool for their project. It is hoped that this thesis has assisted future groups in their activities through a deeper understanding of expectations, possibilities, and results.

8.1 Thesis limitations

During the study, this thesis ran into some difficulties. The first was that the thesis could only review selected tools that expressly fit the practices of one group. Identification of the tools was based on the Nensén project’s goals and transcription model. Other projects will have very different goals and models and therefore tool selection will be different. Nevertheless, this thesis seeks to address this limitation by providing a method for tool selection in place of more concrete recommendations. It is also acknowledged that Uppsala University Library’s implementation of a transcription tool into Alvin will be based on their goals for the cultural heritage database and the expertise of library staff. From the point of view of the Nensén project, a better understanding of transcription as a critical transmission activity has been provided for the libraries involved.

Second, this thesis had planned to look at how the transcribers enacted encoding through XML, but this was shown to be a long-term goal not yet undertaken and could, therefore, not be explored to a greater degree.

Critical assessment of the outcome reveals that the thesis was limited by several factors. Most specifically, that Uppsala has since identified an additional transcription tool, the Virtual Transcription Laboratory (http://wlt.synat.pcss.pl/wlt-web/index.xhtml). This was of interest to Alvin as it supported crowdsourcing (Per Cullhed, personal communication 12 May, 2015). Unfortunately, this tool was not found in the literature review and therefore was not able to be reviewed.
8.2 Future research

It is recommended that future research into transcription tools focus on the Nensén group’s use of markup and how this enables the goals of editorial projects. Many of the tools available make the assumption that digital preservation and analysis of text is sufficiently maintained by simple markup. More practically what is needed is an understanding of these assumptions about tool functionality supporting preservation and accessibility: does markup in this way support these aims?

Specifically, more work could be undertaken to address the roles of the other participants in the Nensén project: namely Umeå University Library and those with system administration roles who would be responsible for implementation of the tool and markup of the text. What is their perspective on this, and how have they decided to approach these issues?

More generally, it could be of interest to look at the concerns of other projects intending to take advantage of transcription tools for presentation in Uppsala’s Alvin database. It is also important that the goals of Uppsala University Library are explored in-depth to discover practical aspects for Alvin that are of consideration for tool application. It would be valuable to gain a better picture of Alvin and the library’s role in this digital archive.
Sources


Appendix I
Interview guide for the J. A. Nensén transcription project group interview

Section 1. Goals

- What are the goals of the Nensén transcription project?
- Will the transcriptions be published? Where?
- How does this project fit into Alvin? What roll with Uppsala University Library play in the project? The other collaborators?
- Why are they being published? Is there current interest in the material? Will publication aim to create interest?
- **Presentation:** How will the published material be presented? As transcriptions only, or with facsimile? What will this include? Text linked to facsimile? Any other presentation features?
- **Use:** How will the material be accessed? What is the digital transcription going to accomplish? Full-text searching, index or database creation, text analysis, public engagement, etc.?
- **Users:** Who are the intended users of the archive? Will the transcriptions allow users to interact with them: make comments or add research? To what degree?
- Is markup necessary? What type of markup? Will it be layered to accommodate future interpretations?
- Will TEI be used? To what extent and why? Will there be a set DTD or created specifically for the project?
- What kind of metadata will be included?
- Other goals? Example: interface functions; link people, places, and dates etc.

Section 2. The Material

- How large is the collection? Where is it located?
- What form do these records (folkminnesuppteckningar) take e.g. diary, notes, correspondence?
- How are the folios laid out? What is the subject content of the records?
- Did he take notes on what he saw/heard or did he also give his opinions/ make comments etc.? Are most of the records his words or collected pieces?
- Handwriting: smooth and neat, but pages are filled to the edges. Are these comments/marginalia or strait text? Are their drawings?
- Language: Swedish, Sami, and Norwegian, 325 pages in Finish, Russian, Latin, Livonian and Polish. How will these affect transcription? Is the language easy for modern readers?
- Are there copyright issues surrounding the material?

Section 3. Technical Considerations

**General:**

- How will the transcription be performed? What programme will be used? Why have you chosen to do the transcriptions the way you have? What are the reasons for this decision?
- Does the tool need to be able to analyse text?
Why has T-Pen been dropped?

Organizational workflow:
- How will the work be approached? How will transcription be performed? In what order will the documents be transcribed, translated?
- How many people are transcribing? How will standardizations be kept? What are they?
- What is the duration of project?
- Will there be a final version, or is the aim to create an on-going edition?

Financial/Technical:
- Does tool cost matter?
- Does tool simplicity matter? Web-based or platform based?
- How will integration of tool into database be achieved?
- Is there staff in place to administrate and programme the software?
- Does licensing matter?

Section 4. Transcription Questions
How the features will be expressed in transcription.

Accidentals (surface features- spelling, punctuation, divisions): graphic evidence
- Letter forms: are there variant forms in 17/18-tals Swedish/Norwegian/Sami? How will these be portrayed?
- Orthography: How will spelling (including misspellings or mistakes?), diacritics (é) be expressed?
- Writing features: Capitalization (as used in the document or normalized/ selected changes?); punctuation (as used in the document/ modernized/ or selected?); Spacing between words, shorthand, symbols, cyphers. Are there any specific examples in Nensén’s records?
- Handwriting: Will each informant be identified- if different handwriting? Will ligatures (&) be shown?
- Abbreviations: How will these be approached- expanded for readability? Will they be marked up?
- Corrections/Emendations: Will the informant’s mistakes/corrections/additions be portrayed/ marked up? If an “obvious” mistake is realised will the transcriber correct/ mark it up?
- Genesis: Will revisions or document development be tracked? Will information in the document about its production be portrayed? Is there any in these documents?
- Surface features: How will the structure and layout be presented? Since the Nensén papers are closely filled with text, will transcription make it more readable or will the effort be to present the text as it appears in the physical document?

Markup: meta-information
- Textuality: Will paragraphs, headings, verses, tables, lists, rubrics, and other structural divisions be marked up?
- Semantics: How will dates, names, places, and keywords be handled?
- Linguistics: Part of speech, lemmatization, syntax?
- Decorations/graphical components: Marginalia, drawings, doodles?
• **Other**: is information about inks, paper, tears, physical dimensions etc. important to the transcription? Will it be included in markup?

**Section 5. Other**

• Is there anything specific to the Nensén records that are of interest for the project to take into consideration when transcribing? For example, references to cultural issues, specific symbols or abbreviation techniques used by Nensén, etc.?

• Any specific goals of the editors themselves: adding commentary, linking to primary sources, cross-referencing, etc.?

Unique characteristics of Nensén’s (1791-1881) records articulated in the grant application to Riksbankens Jubileumsfond (S. Karlgren, personal communication March 18, 2015) which may generate further questions and discussion.

Excerpts of printed material were collected alongside his personal records. **How will these be handled? Why?**

Records stem from a tradition of gathering information on culture and “intangible heritage”. Priests were encouraged to record their observations. Nensén’s records resemble Linnaeus’, but he was also at the forefront of recording the biographic and contextual information on informants (sagespersoner). **Do these traditions have any specific features/characteristics that will be important to convey in transcription? Ex. Language, style, shorthand etc.**

Example of records: 150 pages on Anna Thomadotter who married into the Sami culture and provided many facts about its way of life, including language. Another is first-hand information from Anders Andersson about reindeer and their habits through concrete descriptions. **Has Nensén recorded these in his own words or used the informant’s words? How is this portrayed? Will the transcriptions attempt to keep Nensén’s format and style? Will speakers/writers be identified?**

Unique bibliographic/contextual characteristics: names, dates, places, events.

What are the linguistic/graphic characteristics unique to Nensén: excerpt from records (diplomatic transcription) with resolved abbreviations in brackets:


Problems/concerns about the records: Source validation. Many of the informants could have heard the stories from others (second/third-hand histories).
## Appendix II
Review of the available tools

<table>
<thead>
<tr>
<th>Tool</th>
<th>Developer</th>
<th>Hosted</th>
<th>License</th>
<th>Platform</th>
<th>Cost</th>
<th>Text type</th>
<th>Markup</th>
<th>CMS integration</th>
<th>Document type</th>
<th>Other Features</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABBYY FineReader</td>
<td>ABBYY</td>
<td>Closed Source, Free trial</td>
<td>Windows, Mac</td>
<td>paid</td>
<td>OCR text, various image formats, 190 languages</td>
<td>HTML</td>
<td>Microsoft</td>
<td>OCR engine for editable/searchable documents, PDF</td>
<td>image into text, editing, side-by-side views, tagging</td>
<td><a href="http://finereader.abbyy.com/">http://finereader.abbyy.com/</a></td>
<td></td>
</tr>
<tr>
<td>Alpheios</td>
<td>The Alpheios Project</td>
<td>Open Source, GNU GPL</td>
<td>Web-based</td>
<td>free</td>
<td>Unicoded HTML</td>
<td>TEI, Oxygen editor</td>
<td>Firefox plugin</td>
<td>text</td>
<td>OCR engine for editable/searchable documents, PDF</td>
<td>Read/quiz mode, dictionary, grammar, automatic translation detection</td>
<td><a href="http://alpheios.net/">http://alpheios.net/</a></td>
</tr>
<tr>
<td>Alto Editor</td>
<td>IMPACT Centre of Competence</td>
<td>no</td>
<td>Apache 2.0</td>
<td>Javascript/ Ruby</td>
<td>free-form XML</td>
<td>XML</td>
<td>none</td>
<td>text, image</td>
<td>annotation, tagging, side-by-side image/text box</td>
<td>text, image</td>
<td><a href="http://collatex.net/">http://collatex.net/</a></td>
</tr>
<tr>
<td>CollateX</td>
<td>Interedition</td>
<td>GNU, GPL 3</td>
<td>Windows, Mac, Linux, Android, iOS</td>
<td>free</td>
<td>Unicode plain text, JSON, XML</td>
<td>TEI, XML</td>
<td>text</td>
<td>collation, Java software</td>
<td>visualization</td>
<td><a href="http://collatex.net/">http://collatex.net/</a></td>
<td></td>
</tr>
<tr>
<td>digilib</td>
<td>Max Planck Institute for the History of Science, University of Bern</td>
<td>yes</td>
<td>Open Source, GNU GPL</td>
<td>free</td>
<td>various image formats</td>
<td>Javascript plugin, URLs</td>
<td>images</td>
<td>annotation, stable URLs</td>
<td>zoom, rotation, mirrored, colour enhancement</td>
<td><a href="http://digilib.sourceforge.net/">http://digilib.sourceforge.net/</a></td>
<td></td>
</tr>
<tr>
<td>Tool</td>
<td>Developer</td>
<td>Hosted</td>
<td>License</td>
<td>Platform</td>
<td>Cost</td>
<td>Text type</td>
<td>Markup</td>
<td>CMS integration</td>
<td>Document type</td>
<td>Other</td>
<td>Interface features</td>
</tr>
<tr>
<td>----------------------</td>
<td>-------------------------------</td>
<td>--------</td>
<td>---------------------</td>
<td>-------------------</td>
<td>------</td>
<td>-----------</td>
<td>--------</td>
<td>-----------------</td>
<td>---------------</td>
<td>------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Digital Mappaemundi</td>
<td>Martin Foys, Shannon Bradshaw</td>
<td></td>
<td></td>
<td>Web-based</td>
<td>free</td>
<td>Structured</td>
<td>XML</td>
<td>text, image, fragments</td>
<td></td>
<td>text, image, commentary</td>
<td>annotation and linking of text and images, supports different editors, commentary</td>
</tr>
<tr>
<td>ediarum</td>
<td>Stefan Dumont, Martin Fechner</td>
<td></td>
<td></td>
<td>Windows, Mac, Linux</td>
<td>other</td>
<td>TEI, XML</td>
<td>TEI, Oxygen editor</td>
<td>Java</td>
<td>text</td>
<td>text, image, comment</td>
<td>TEI toolbar</td>
</tr>
<tr>
<td>eLaborate</td>
<td>Huygens Instituut KNAW</td>
<td>yes</td>
<td>GNU GPL3</td>
<td>Windows, Mac, Linux</td>
<td>free</td>
<td>free-form</td>
<td></td>
<td>text, image</td>
<td></td>
<td>text, image, annotation, publish</td>
<td>diplomatic and text layers</td>
</tr>
<tr>
<td>FromThePage</td>
<td>Ben Brumfield</td>
<td>yes</td>
<td>Open Source, Affero GPL</td>
<td>Web-based, Ruby on Rails</td>
<td>free</td>
<td>free-form</td>
<td>Wiki markup, TEI as export</td>
<td>Archive.org, Omeka</td>
<td>text</td>
<td>text, image, annotation, publish</td>
<td>collaborative semantic markup for indexing, annotation</td>
</tr>
<tr>
<td>Islandora</td>
<td>Robertson Library</td>
<td>no</td>
<td>GPL 3.0</td>
<td>Drupal, Fedora</td>
<td>free</td>
<td>free-form</td>
<td>TEI</td>
<td>text, image, video</td>
<td></td>
<td>text, image, annotation, publish</td>
<td>organisie manage</td>
</tr>
<tr>
<td>Omeka</td>
<td>Roy Rosenzweig Center for History and New Media</td>
<td>yes</td>
<td>Open Source, GNU GPL</td>
<td>Web-based</td>
<td>free</td>
<td>various forms</td>
<td></td>
<td>text, image</td>
<td></td>
<td>text, image, annotation, publish</td>
<td>collaborative archive collection development choice of themes</td>
</tr>
<tr>
<td>Proofread Page</td>
<td>ThomasV, Tpt</td>
<td></td>
<td>GNU GPL 2</td>
<td>Web-based, MediWiki</td>
<td>free</td>
<td>free form, facsimies</td>
<td>PHP</td>
<td>text, image</td>
<td></td>
<td>OCR, early print/ manuscript</td>
<td>text box</td>
</tr>
<tr>
<td>Tool</td>
<td>Developer</td>
<td>Hosted</td>
<td>License</td>
<td>Platform</td>
<td>Cost</td>
<td>Text type</td>
<td>Markup</td>
<td>CMS integration</td>
<td>Document type</td>
<td>Other</td>
<td>Interface features</td>
</tr>
<tr>
<td>---------------</td>
<td>--------------------------------</td>
<td>--------</td>
<td>-------------------------------</td>
<td>-------------------------------</td>
<td>--------</td>
<td>-----------</td>
<td>--------</td>
<td>-----------------</td>
<td>--------------</td>
<td>------------------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>PyBossa</td>
<td>Citizen Cyberscience Centre/DKFN</td>
<td>yes</td>
<td>Open Source, GNU Afero GPL</td>
<td>Python, Gdoc</td>
<td>free</td>
<td>Tabular, free form</td>
<td>no TEI</td>
<td>Python</td>
<td>text</td>
<td>collaborative, annotation</td>
<td>data link to images, crowdsourcing</td>
</tr>
<tr>
<td>Scribe</td>
<td>Zooniverse</td>
<td>optional</td>
<td>MIT, ASL</td>
<td>jQuery, Ruby on Rails</td>
<td>free</td>
<td>structured data</td>
<td>no TEI</td>
<td>none</td>
<td>text, image</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scripto</td>
<td>Center for History and New Media at George Mason University</td>
<td>no</td>
<td>GPL 3.0</td>
<td>PHP library, MediaWiki</td>
<td>free</td>
<td>free-form, wikitext</td>
<td>Wiki markup</td>
<td>Omeka, WordPress, Drupal</td>
<td>text, image</td>
<td>collaborative, diplomatic</td>
<td></td>
</tr>
<tr>
<td>TEI Boilerplate</td>
<td>Indiana University</td>
<td>yes</td>
<td>Open Source, Creative Commons Attribution</td>
<td>Windows, Mac, Linux</td>
<td>free</td>
<td>free-form, facsimiles</td>
<td>XML, TEI P</td>
<td>Omeka</td>
<td>text</td>
<td>served to web without server-side processing or HTML.</td>
<td>choice of document styles</td>
</tr>
<tr>
<td>Tesseract</td>
<td>Ray Smith, HP Labs, Google</td>
<td>Open Source, Apache</td>
<td>Windows, Mac, Linux</td>
<td>Various image formats, 60 languages</td>
<td>free</td>
<td>OCR conversion of images</td>
<td>text, image</td>
<td></td>
<td></td>
<td><a href="https://code.google.com/p/tesseract-ocr/">https://code.google.com/p/tesseract-ocr/</a></td>
<td></td>
</tr>
<tr>
<td>T-Pen</td>
<td>St. Louis University Center for Digital Theology</td>
<td>Open Source, Educational Community License (EPL 2.0)</td>
<td>Web-based, Java/JavaScript</td>
<td>free</td>
<td>free-form, XML</td>
<td>TEI, XML</td>
<td>none</td>
<td>text, image</td>
<td>collaborative editing, automatic parsing, manual parsing, line by line transcription, comments</td>
<td></td>
<td><a href="http://t-pen.org/TPEN/">http://t-pen.org/TPEN/</a></td>
</tr>
<tr>
<td>Tool</td>
<td>Developer</td>
<td>Hosted</td>
<td>License</td>
<td>Platform</td>
<td>Cost</td>
<td>Text type</td>
<td>Markup</td>
<td>CMS integration</td>
<td>Document type</td>
<td>Other</td>
<td>Interface features</td>
</tr>
<tr>
<td>----------------------</td>
<td>---------------------------------</td>
<td>--------</td>
<td>------------------</td>
<td>-------------------</td>
<td>-------</td>
<td>-----------</td>
<td>--------</td>
<td>-----------------</td>
<td>---------------</td>
<td>-----------------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>Transcription Assistant</td>
<td>Worcester Polytechnic Institute</td>
<td>no</td>
<td></td>
<td>Windows, Mac</td>
<td>free</td>
<td>free-form</td>
<td>MML</td>
<td>text, image</td>
<td>metadata for search and archive</td>
<td>line detection</td>
<td><a href="http://www.wpi.edu/Pubs/E-project/Available/E-project-042507-120230/unrestricted/Web-Based_Emergent_Manuscript_Transcriptions_D07.pdf">http://www.wpi.edu/Pubs/E-project/Available/E-project-042507-120230/unrestricted/Web-Based_Emergent_Manuscript_Transcriptions_D07.pdf</a></td>
</tr>
<tr>
<td>Transcription</td>
<td>University of London Computer Centre; UCL Bentham Project</td>
<td>yes</td>
<td>GPL 2.0</td>
<td>MediaWiki</td>
<td>free-form</td>
<td>Simple TEI</td>
<td>text</td>
<td>structures data, edit, organize</td>
<td>collaborative</td>
<td>TEI toolbar, preview</td>
<td><a href="http://www.transcribe-bentham.da.ulcc.ac.uk/st/Transcribe_Bentham">http://www.transcribe-bentham.da.ulcc.ac.uk/st/Transcribe_Bentham</a></td>
</tr>
<tr>
<td>Transcrivia</td>
<td>Bartas Technologies</td>
<td></td>
<td>Transcrivia license</td>
<td>Mac</td>
<td>paid</td>
<td>free form</td>
<td>text, audio, video</td>
<td>optional TEL</td>
<td>correction of text-version of page images</td>
<td>recording, play back</td>
<td><a href="http://www.bartastechnologies.com/products/transcrivia/">http://www.bartastechnologies.com/products/transcrivia/</a></td>
</tr>
<tr>
<td>TypeWrite</td>
<td>18th Connect</td>
<td></td>
<td>Web-based</td>
<td>free</td>
<td>free-form</td>
<td>optional TEL</td>
<td>text, image</td>
<td>TEI, METS, ALTO</td>
<td>navigation section, text correction section</td>
<td><a href="http://www.18thconnect.org/typewright/documents">http://www.18thconnect.org/typewright/documents</a></td>
<td></td>
</tr>
<tr>
<td>Veridian</td>
<td>DL Consulting</td>
<td>yes</td>
<td>Proprietary</td>
<td>Veridian</td>
<td>paid</td>
<td>free-form, structured</td>
<td>none</td>
<td>text, image</td>
<td>collaborative, OCR</td>
<td>designed for large-scale projects</td>
<td><a href="http://www.veridiansoftware.com/">http://www.veridiansoftware.com/</a></td>
</tr>
<tr>
<td>Xalan</td>
<td>Apache Software Foundation</td>
<td></td>
<td>Apache, Open Source</td>
<td>Windows, Mac, Linux</td>
<td>free</td>
<td>structured data, stylesheets</td>
<td>Java</td>
<td>text</td>
<td></td>
<td></td>
<td><a href="http://xalan.apache.org/">http://xalan.apache.org/</a></td>
</tr>
</tbody>
</table>