Legacy Data in a Digital Age

Ellert Thor Johannsson, Simonetta Battista and Tarrin Wills [0000-0001-5360-3495]

[1]A Dictionary of Old Norse Prose, dept. of Nordic Studies, University of Copenhagen
ellert@hum.ku.dk, sb@hum.ku.dk, tarrin@hum.ku.dk

Abstract. In this article we discuss the evolution of data collection and use in the context of *A Dictionary of Old Norse Prose* (ONP), which is a dictionary covering the medieval language of Iceland and Norway. This dictionary started as a collection of citations on paper slips, which were later used as the lexicographical material for a multi-volume, but incomplete, print edition. Today the citations form the basis of the digital version of ONP, currently available online at onp.ku.dk [1]. We account for the evolution of the data within three different periods, separated by certain milestones. We demonstrate how the legacy data have evolved and been enhanced throughout the years to provide innovative ways to bring together and take advantage of all the information gathered by ONP during its existence.

Keywords: Lexicography, Online Dictionary, Databases.

1 Introduction

A Dictionary of Old Norse Prose (ONP) is a historical dictionary, which covers the language of Iceland and Norway in the Middle Ages. This dictionary began as an extensive collection of citations on paper slips, which were excerpted from all known works of Old Norse Prose texts. Later developments include a multi-volume, but incomplete, print publication and an online lexicographic tool [1], providing detailed information about the vocabulary of Old Norse and its textual foundation in medieval manuscripts and documents.

Even though the semantic analysis of the vocabulary is not yet fully completed, the wide scope of the dictionary is evident by the fact that its archive of around 800.000 example citations represents around 7% of the estimated 11 million word corpus of Old Norse prose from known sources. The long history of the project provides an opportunity to study the development of the data and how they have been used during the time the project has been in existence.

The goal of this article is to account for the legacy data of ONP and how they have been acquired, organized and utilized. For that purpose, we divide the history of the dictionary into three periods that roughly coincide with the three stages of its development. The first period extends from the foundation of the dictionary to the making of its first database; the second period covers the early computer assisted lexicographic

work until the establishment of the online version of the dictionary and the third period spans from the launch of the online dictionary until present times.

2 1st Period – The Early Days

Work on this dictionary began in 1939 long before computers and databases became available. The nature of the material and the editorial principles set out by the founders of the project demanded a wide variety of data be collected and organized [2]. This primarily involved excerpting the source material for representative examples of word use.

2.1 The Data

The essence of the dictionary is the collection of example citations which today number more than 800.000, each of which is provided with a sentence illustrating a specific form and/or meaning of the headword, a detailed reference showing the work of origin as well as page and line number.

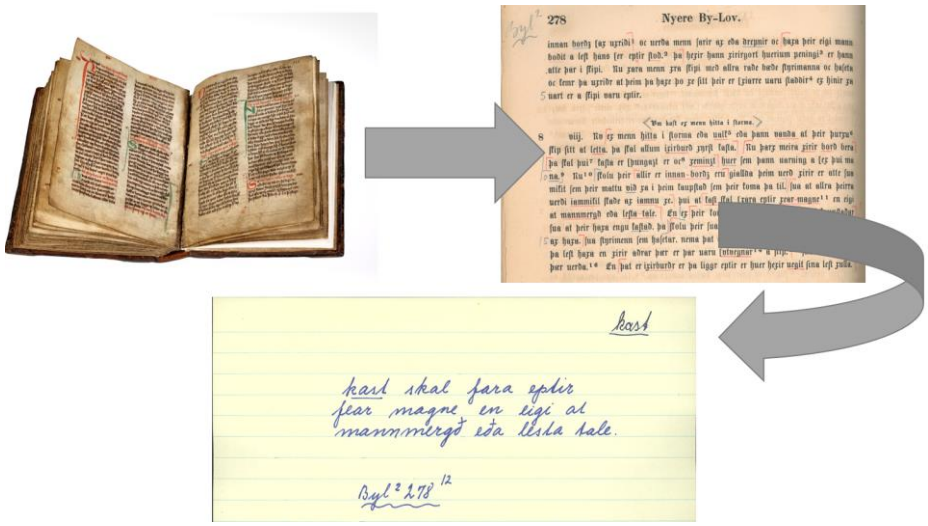

Fig. 1. The medieval manuscript is edited in a scholarly edition, which in turn is excerpted by underlining relevant citations and writing them down on paper slips.

Most citations came into being in the early days when the dictionary staff would read scholarly text editions in order to collect significant examples of word use by underlining sentences that contained a lemma in its syntactic context (cf. Fig 1). Once the examples had been selected, a copyist would write down each example sentence on a paper slip with the relevant information. In some cases where a scholarly edition was lacking, words were excerpted directly from the manuscript sources. Around 9% of the

total number of excerpted citations contain a direct reference to a manuscript. Additional information was also noted on the slips, e.g. variant manuscript readings, textual commentary or corresponding words from foreign sources. The slips were then filed in a physical archive. Further detail about the slips and the information they contain can be found in [3].

Besides the citations, it was important to record various information relating to the source material. Structuring these data involved creating an index of all the different medieval works, which had been excerpted. The citations included a reference to scholarly editions as well as the manuscripts these were based on, and all this information had to be registered. In addition to this, various other data were gathered about the medieval texts, such as the dating of manuscripts, bibliographic details about editions and secondary literature, as well as other supplementary information. In the early days this work was also registered on paper through various filing systems.

3 2nd Period – The Database and Print Publication

In the 1980s, the potential advantages of working with the data in a database structure became clear and work began on developing the necessary digital tools to process the dictionary material. The ideal database needed to accommodate all the dictionary citations as well as the data related to the source material. An evaluation report of the project from 1993 gives an insight into the thought process and considerations behind the design of the database [4].

After experimenting with various options, an Oracle-based SQL database became the IT solution of choice for bringing together the complex data ONP had gathered throughout the years. The interconnectivity of the data demanded the creation of many different database tables, and sets of tables, where all the pieces of information were stored in separate fields.

In total, the database included more than 200 tables that were used in varying degrees in the dictionary making process. The most important ones are the wordlist table, headword table, definition table, citation table, as well as tables for references to editions, secondary literature and other glossaries. The tables were linked together through common fields. This allowed supplying information relating to both the source material and each citation, e.g. the notation of the geographical provenience of the manuscripts and the grouping of the source material by literary genres.

The benefits of organizing the information in a database were immediate. Once the content from the hand-written index registry had been entered into the database a printed index volume could be published in 1989 [5]. Even though this volume is primarily conceived to facilitate the use of the then unpublished dictionary, it stands alone as an independent reference work covering Old Norse prose texts and their manuscript origins.

There were further challenges when introducing the new technology because of the nature of the material, especially the widespread use of non-standardized characters and symbols. In many cases, the system developer had to combine standard symbols to cre-

ate special characters. This led to the development of a tailor-made ONP font. Eventually the dictionary implemented sufficient technical resources to facilitate the structuring of dictionary entries for each of the 65.000 headwords and begin the work of publishing them.

3.1 Print Dictionary

The organizing of information in a database was undertaken with the end goal in mind to publish the dictionary entries in traditional printed volumes. Once the citations had been keyed into the database the editors could proceed with the structuring of dictionary entries, writing definitions, supplying extra grammatical information, as well as information about collocations and syntactic relations. This work resulted in an ambitious publishing plan of 12 dictionary volumes along with the aforementioned index volume. Three printed volumes were eventually published from 1995 to 2004, containing entries that cover the alphabet from *a-em* [6]. Even though the print publication was well received, it had its limitations. It took about five years to prepare each volume and space restrictions meant that many citations remained unpublished.

4 3rd Period – Online Dictionary

After the publication of the third volume, the print edition was put on hold. The project underwent another restructuring process, which resulted in an online version made available in 2010. An important step in the conversion of the paper dictionary into a full-fledged digital online dictionary was the scanning of ca. 500.000 hand-written paper slips, which were integrated into the database and linked to the same fields as the typed citations [3]. The digitalization process also included scanning scholarly editions to provide extended context for each example.

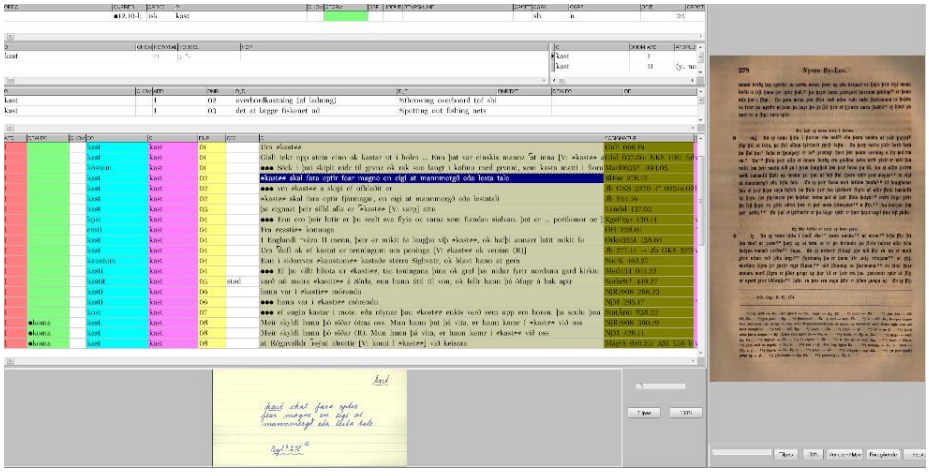

Fig. 2. The editing interface after digitalization, with five database tables, as well as a scanned slip for each citation and a scanned page from a relevant text edition.

After this restructuring, the database could be used efficiently in the editorial work with all the information being available to the dictionary staff. The new editing interface provided access to all the legacy data as well as additional information about each citation and each medieval source text (cf. Fig 2).

The database also became an essential part of the published dictionary as online users could query the database directly and search the data in different ways, being no longer limited by the alphabetical order of printed dictionary entries. Besides headword search, the database structure made it possible to tailor the search to particular parts of speech or lexical items [7]. Scanned slips and editions proved also to be very useful for online users as they now could access the dictionary information even in cases where the headwords had not been edited yet.

4.1 Enhancing the Data and Web Applications

Since 2010 the online version of ONP has been gradually expanded with new edited entries and improved with additional search capabilities, such as text provenance and literary genre. Moreover, the dictionary data have been enhanced by linking them internally and to other digital resources. These expanded functions constitute an important part of the redesigned website and user interface, which was launched in 2019.

The prominent components of this new website are two new web applications: an integrated web publishing and editing application, and a fast, archivable public interface. Both applications use the dictionary's Oracle RDBMS back-end and an interface using PHP that interacts with the database to generate HTML and/or JSON output. The applications use Bootstrap as the HTML framework and user interaction is coded in JavaScript. These new applications take advantage of SQL in joining multiple tables in complex queries and allow for linking in various ways between the semantic tree and citations, pages and headwords [8]. A discussion of the website functions along with a description of the data structure and an overview of technical features is also found in [8].

The applications allow the user to interact with the data in many different ways as well as to access other dictionaries and resources (cf. Fig. 3), including digital editions of Old Norse texts and manuscript images [9].

Fig. 3. ONP Online provides references to older dictionary works in association with each entry. This screenshot shows information from two older dictionaries.

One important innovation is the “ONP Reader”, which provides the user with glossaries to scholarly text editions (cf. Fig. 4). This allows the user to access all the words excerpted from each page of a particular text [8]. Since ONP tends to excerpt many texts in great detail, the “ONP Reader” has expanded the dictionary’s potential user base to students and others less familiar with non-standard Old Norse texts. With all its extra features and easy access to secondary supplementary information, ONP stands as an even more important research tool for scholars in medieval Scandinavian language, literature, and culture.

Fig. 4. An example page displayed in the “ONP Reader” of ONP Online. A selected section of words from line 12 has been enlarged here to facilitate reading.

5 Conclusions

In this article, we have shown how legacy data of the ONP dictionary, originally only organized in paper filing systems, have been structured in a database and improved in various ways throughout the project’s history. The original data still provide the basis of the lexicographic work, and constitute a unique collection of representative examples. The database structure and linking of information has facilitated the use of these data in many different ways, and made them accessible to users from all over the world, long before the editing of the dictionary has been completed. Search capabilities made possible in the online digital version of the dictionary and later linking of the data to

external sources further enhance the value of the legacy data and provide innovative ways to bring together and take advantage of all the information gathered by ONP during its existence.

References

1. ONP Online = Ordbog over det norrøne prosasprog / A Dictionary of Old Norse Prose Online. Accessed at: http://onp.ku.dk, last accessed 2020/10/10
2. Widding, O.: Den Arnamagnæanske Kommissions Ordbog, 1939–1964: Rapport og plan, G.E.C. GADS Forlag, Copenhagen (1964).
3. Johannsson, E.: "Integrating analog citations into an online dictionary". In: Navarretta, C., Agirrezabal, M., Maegaard, B. (eds.). Proceedings of the Digital Humanities in the Nordic Countries 4th Conference, pp. 250–258. RWTH, Aachen (2019).
4. ONP 1993 = Evaluation of the Production Plan for the Dictionary of Old Norse Prose. Ministry of Education and Research, Copenhagen (1993).
5. ONP Registre = Degnbol, H., Jacobsen, B. C., Rode, E., Sanders, C. & Helgadóttir, Þ. (eds.). Ordbog over det norrøne prosasprog / A Dictionary of Old Norse Prose. Registre/Indices. Den Arnamagnæanske Kommission, Copenhagen (1989).
6. ONP 1-3 = Degnbol, H., Jacobsen, B. C., Knirk, J. E., Rode, E., Sanders, C. & Helgadóttir, Þ. (eds.). Ordbog over det norrøne prosasprog / A Dictionary of Old Norse Prose. ONP 1: a-bam (1994). ONP 2: ban-da (2000). ONP 3: de-em (2004). Den Arnamagnæanske Kommission, Copenhagen.
7. Johannsson, E., Battista, S. "Editing and Presenting Complex Source Material in an Online Dictionary: The Case of ONP". In: Margalitadze, T., Meladze, G. (eds.). Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity, 6–10 September 2016, pp. 117–128. Ivane Javakhishvili Tbilisi State University, Tbilisi (2016).
8. Wills, T., Johannsson E.: "Reengineering an Online Historical Dictionary for Readers of Specific Texts". In: Kosem, I., Zingano Kuhn, T., Correia, M., Ferreria, J. P., Jansen, M., Pereira, I., Kallas, J., Jakubíček, M., Krek, S. & Tiberius, C. (eds.). Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference. 1–3 October 2019, Sintra, Portugal, pp. 116–129. Lexical Computing CZ, s.r.o, Brno (2019).
9. Wills, T., Johannsson, E., Battista, S.: "Linking Corpus Data to an Excerpt-based Historical Dictionary". In: Čibej, J., Gorjanc, V., Kosem, I., Krek, S. (eds.). Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, pp. 979–987. Ljubljana University Press, Faculty of Arts, Ljubljana (2018).