Digitization of the Collections at Ømålsordbogen – the Dictionary of Danish Insular Dialects: Challenges and Opportunities

Henrik Hovmark
Asgerd Gudiksen
Department of Nordic Studies and Linguistics, University of Copenhagen, Denmark

1 Summary

Ømålsordbogen (the Dictionary of Danish Insular Dialects, henceforth DID) is an historical dictionary giving thorough descriptions of the dialects, i.e. the spoken vernacular of peasants and fishermen, on the Danish isles Seeland, Funen and surrounding islands. It covers the period from 1750 to 1950, the core period being 1850 to 1920. Publishing began in 1992 and the latest volume (11, *kurv-lindorm*) appeared in 2013 but the project was initiated in 1909 and data collection dates back to the 1920s and 1930s. The project is currently undergoing an extensive process of digitization: old, outdated editing tools have been replaced with modern (database, xml, Unicode), and the old, printed volumes have been extracted to xml as well and are now searchable as a single xml file. Furthermore, the underlying physical data collections are being digitized.

In the following we give a brief account of the latter digitization process, involving the physical collections, and we discuss a number of questions and dilemmas that this process gives rise to. The collections underlying the DID project comprise a variety of sub-collections characterized by a large heterogeneity in terms of form as well as content. The information on the paper slips is usually densified, often idiosyncratic, and normally complicated to decode, even for other specialists. The digitization process naturally points towards web publication of the collections, either alone or in combination with the edited data, but it also gives rise to a number of questions. The current digitization process being very basic, only very few metadata (1-2 or 3) can be added during the scanning process, we point to the obvious fact that web publication of the collections presupposes an addition of further, carefully selected metadata, taking different user needs and qualifications into account. We also discuss the relationship between edited and non-edited data in a publication perspective. Some of the paper slips are very difficult to decipher due to handwriting or idiosyncratic densification and we point out that web publication in a raw, i.e. non-edited or non-annotated form, might be more misleading than helpful for a number of users.

2 Dictionary and Collections as Cultural Heritage

The DID project and the underlying collections of data are an important part of Danish cultural heritage and cultural preservation (DID is not exceptional in this respect, cf. for instance Grønvik 2016). First, the collections and DID contain unique information about Danish language, not only the spoken vernaculars but also Danish language in an historical context. Second, DID gives thorough descriptions of the culture and life world of the dialect-speaking peasants and fishermen along with the detailed linguistic information about pronunciation, morphology, syntax and semantics. When the first specific plan for DID was outlined in the 1920s, the project was inspired by the German Wörter und Sachen tradition that stressed the importance of the cultural context in which words are used, and DID was defined as a so-called "sproglig-saglig ordbog" ('linguistic-cultural dictionary', cf. the subtitle of DID). Thus, the collections and DID contain systematic information about folklore, traditions, feasts, etc. (*fastelavn* 'Shrovetide') as well as descriptions

of (parts of) tools (*le* 'skythe', *kærne* 'churn', etc.), work processes (*høst* 'harvest', *bagning* 'baking', etc.), central crops (*hør* 'flax', *kartoffel* 'potato', etc.) and artifacts (*træsko* 'clogs', etc.). In DID, this is reflected not only in the general selection of lemmas but also in separate encyclopedic parts of relevant articles, sometimes as lengthy as 3 to 4 columns (cf. the entries *høst* 'harvest' or *bage* 'bake').

3 The Collections: Form and Characteristics

The collections describing the traditional dialects at the Centre for Dialectology comprise a substantial physical collection (mainly paper slips, around 3.5 million). The collections also comprise a large number of recordings, mainly tape recordings from the 1970s but older as well as newer recordings exist. A minor part of the recordings has been transcribed and is now used as a small corpus containing highly valuable information about linguistic aspects that are less well represented in the paper slip collection, for instance function words, syntax, and pragmatic particles. In the following we focus on the physical collection, which can be divided into four functionally organized subparts, each containing a number of sub-collections.

3.1 The Edited Alphabetic Collection

During the editing process all paper slips from the different sub-collections are brought together in one, single alphabetic collection. The organization of the paper slips under each entry word is identical to the internal structure of the entries in the dictionary. This collection contains around half of the total paper slips, i.e. the data behind volumes 1-11 (*a-lindorm*; now also available as a roughly tagged xml file) as well as the forthcoming volume 12 (*lindost-mas*…; also available as an xml database with a large amount of specialized metadata).

3.2 The Non-edited Alphabetic Collections

The collections of paper slips etc. that have not been edited yet are characterized by a very complex organization. They are divided into several alphabetically ordered sub-collections, first of all 50 collections subdivided according to topic (for instance "harvest" or "baking", cf. section 2). To this should be added a large collection of paper slips with general, i.e. non-topic related information, as well as an even larger collection of paper slips with accession data.

3.3 Background and Source Collections

A number of collections can be described as background or source collections. They contain first of all a topographically ordered collection and a manuscript collection. The topographic collection contains various sources and records, often longer descriptions giving information about the contextual use of vocabulary but also transcripts of interviews as well as data from specialized questionnaires etc. To this should be added around 20 alphabetically ordered sub-collections of pronunciation and phonology from different parts of the Danish Isles as well as a collection of corresponding sub-collections from Jutland and Bornholm (i.e. the East Danish dialect area). A photo collection, primarily containing photos and drawings of tools and artefacts (cf. the 50 topic collections), also belongs to this group.

3.4 Specialized Editor Collections

The specialized editor collections contain different internal reference tools (archive of informants, manuscripts, sources, topic organization etc.).

4 Dilemmas and Challenges

Initially part of an independent, private institution (*Udvalg for Folkemål*, The Dialect Commission), the project became part of the University of Copenhagen in 1960 (*Institut for Dansk Dialektforskning*, the Institute of Danish Dialectology). Thus, scholar institution and heritage institution are in this case identical, i.e. the University of Copenhagen. This does not mean, however, that the digitization process is without challenges. First, universities are in general first of all defined as institutions for education and research and not as institutions taking care of cultural heritage (exceptions do exist, cf. for instance the Medical Museion at the University of Copenhagen – www.museion.ku.dk). Second, neither editor nor scholar is necessarily used to think in terms of digitization and communication on new platforms. Especially it is a challenge to make a realistic integration of future users in a digitization strategy which has web publication as its primary goal. During the present digitization process we were asked to make a prioritized list of the different parts of the physical collections, which made us consider different criteria.

4.1 The Digitization Project – Overview, Plan and Scope

The digitization project started around November 2016 and receives funding until June 30th 2018. It was initiated as part of a harsh cost cut plan at the University of Copenhagen and the Faculty of Humanities. Budget expenses being partly based on amount of square meters used by the department, it was decided to launch a project of digitization of all the physical collections at the department of Nordic Research (now part of the Department of Nordic Studies and Linguistics) in order to vacate two entire floors and gain new income from renting out these floors to external lease-holders. The physical collections will be moved to a storage room in the basement, and future access and use are supposed to be done on the basis of the digitalized files (unless special consultation of the physical document is needed). Thus, the digitization project is closely defined to this practical and urgent end. The unique collections at the section for Name Research (including for instance old maps of various sizes) as well as various documents at the Arnamagnæan Institute (mainly a photo collection, the medieval hand-written manuscripts being stored in a special storage room) are also being digitized as part of the project.

Technically, the digitization is done using hardware from Fujitsu and the scanning software Kofax Capture. Each scan is saved as a separate .tiff-file with a unique reference number. The .tiff-files are currently being converted to parallel .jpg-files in order to facilitate future browsing and download. Each, single paper slip or sheet is treated as a separate scanned document. Kofax Capture enables the addition of metadata during the scanning process. However, the software is not ideal for adding numerous metadata dynamically, and such a process would also require specialist involvement and time-consuming analysis. Therefore, only very few metadata are added during the very scanning process due to the fact that the main, overall aim of the project is to scan as many documents as possible. Instead, it is the plan to locally develop a browsing tool as well as tailor-made metadata software adapted to the needs of each local project (for instance DID). This process has only been initiated.

The digitization process will entail a number of advantageous results: digital back-up of all the collections, safer storage in a special storage room, and a much broader possible use of the collections in the future. However, the process is currently dependent on a successful development of browsing and tagging tools. We would also like to point out that a successful digitization, i.e. a digitization that results in a broader use of the collections, requires further metadata, and, not least, considerations of future users and user needs. The addition of further metadata is probably beyond the scope of the current digitization

project, financially as well as practically, and separate funding will be needed. We discuss these points in the following in relation to the specific characteristics of the DID project.

4.2 Internal Use of the Collections

Users are in general thought of as external readers: in the case of DID, for instance, the typical readers have traditionally been defined as fellow researchers within linguistics, ethnology or folklore, or local historians with a special interest in dialects or local history. But given the fact that the collections were established with a future dictionary in mind, the paper slips also contain a large amount of internal information, and, consequently, the users are also the editors at the DID project itself. This applies obviously to the internal sub-collections (cf. section 3.4) but also to the other sub-collections, especially the non-edited paper slips (cf. section 3.2.). The sub-collections with high relevance for the editors have been given high priority in the digitization process due to the importance of continuous productivity and maintenance of the professional standards and competence at the project. The challenge here is primarily to develop digital editing tools that are sufficiently efficient and dynamic: developing tools that will enable simple consultation of information in an alphabetically ordered (sub)collection is rather straightforward but it is still a question if it is possible to develop an online editing tool which is able to handle the very complex process involved in analyzing and (re)sorting sometimes numerous paper slips, including the addition of metadata (big articles may be based on hundreds of paper slips and contain more than 20 meanings) (cf. Bakker & Grønvik (2008) on the development and use of such a tool at the Norwegian dictionary Norsk Ordbok, which has a many similarities with the DID project).

The development of the editing tools has only been initiated (cf. section 4.1) and the final result is still unknown. Given these somewhat unclear circumstances, we have prioritized the development of an efficient browsing tool, which is a prerequisite for the necessary continuous access to and use of the collections by internal as well as external users. However, we are also focusing on the possibility of adding further metadata to each scanned document. We are considering which metadata are most central to the collections and most relevant for future users and uses. We discuss this in the remaining two sections.

4.3 Web Publication, New External Users and the Need for Metadata

It is evident that the digitization process also has to take external users into account. Furthermore, web publication will soon be relevant, and the number and nature of external users are expected to be larger than before; thus, the qualifications of future users will be more diverse and less specialized. This development is very welcome indeed but it also has implications for the priorities to make.

DID and the collections contain a large amount of information with appeal to a larger public. This goes for the information about the traditional peasant culture (for instance work processes or feasts, games or folklore) as well as the very diverse vocabulary attested (for instance nicknames). All information in DID as well as in the collections is also localized geographically, usually or if possible to a parish, i.e. a comparably small geographical entity. The geographical metadata also have a large potential in terms of web publication, not only technically (GPS, GIS) but also because of an increasing societal interest in local meaning and history in a globalized world.

The linguistic, ethnological and geographical information is thoroughly described and therefore quite easily accessible for a wider public in the edited dictionary. The descriptions in the topographic collection, which are often coherent and longer, are also comparably easy to decode and, thus, suited for publications on the web. Consequently, these two collections received the highest priority in the

digitization process: we might imagine, for instance, a web publication of the printed volumes with the possibility of looking at the paper slips used for each entry, links to photos in the photo collection and the possibility of creating maps showing where a word, meaning or pronunciation is attested in the data; or publication of the topographic collection enriched with geographical links and ethnographical metadata (cf. the subdivision into 50 topics). The manuscript collection also contains a large number of very interesting texts that are suited for web publication, although for a more specialized public (old handwritten manuscripts of relevance for linguists or historians, or newer, unpublished theses).

Consequently, web publication of a number of the sub-collections has a large potential and we have prioritized in accordance with this perspective. This does not mean, however, that the sub-collections are ready for publication at present. The main challenge is, again, the lack of metadata. In order to make the information in the most relevant sub-collections accessible and, thus, realize the rich publication potentials for a wider public, further addition of salient metadata is required.

During the present initial work with a browser and editing tool, we chose to distinguish between an initial tagging phase and an editing phase. In the tagging phase each paper slip goes through an additional tagging process where further metadata can be added; in the editing phase, detailed sorting(s) and analysis of a whole set of paper slips would take place. When we started sketching the tool in cooperation with the IT department, our primary focus was on the editing process and the metadata were chosen according to their relevance for the subsequent editorial process. Currently, however, we are going through the list of metadata again, reconsidering which metadata would (also) be useful in a future web publication. Information about headword (and word class) is added during the scanning process but in addition to this, we are giving priority to information about geographical location and ethnological-folkloristic topic. We also consider information about source, i.e. informant, collection and/or literary source, to be relevant (cf. section 3.3). Information about the record process and situation might be relevant too, for instance the record year, however not necessarily any detailed information, for instance about the researcher who conducted or transcribed an interview.

The before-mentioned metadata are characteristic of the collections as a whole, and a future link between sub-collections (for instance the photo collection, the topographical collection and/or the alphabetic collections) could be anticipated. A future link between the edited volumes and the paper slips could also be anticipated. We see the tagging of the scanned documents with these metadata as a way of consolidating the coherence between the different parts of the collection in a future digital environment, and also facilitating future web publication. However, any web publication would still require considerations about specific intended users, their needs and qualifications and the addition of further, more specialized metadata would possibly be necessary. Seen as a whole, the collections are quite complex and contain a wide variety of knowledge, which often requires specialized expertise in order to be fully understood (cf. Engerer et al. 2017).

4.4 Publication of Raw, Non-edited Data?

The digitization makes it possible to publish the raw, non-edited data directly on the web, along with edited data but also alone. A number of the sometimes abundant informations on paper slips pertain to the salient and easily recognizable characteristics of the DID project, for instance geographical location, and other kinds of information are standard in kind, for instance information about sources, informants or word class. A great deal of the information, however, is not standard but very specialized or even idiosyncratic (cf. Tasovac & Petrović 2015). Many paper slips contain internal messages to other editors, often in

abbreviated forms or making reference to internal archives or even editing rules that do not exist in any written form. The abbreviation “lb.nr.”, for instance, is short for ‘serial number’ and refers to a specific section in a special questionnaire organized according to the central ethnological topics (cf. section 2). As such, it is actually quite important information but it is delivered in an obscure form. Also, standard information is regularly given in unorthodox or incomplete form, for instance information about geographical location or informants. Other kinds of information require specialized scholarly guidance. The phonological notation, for instance, displays a huge variation on the paper slips, and some variations are idiosyncratic (due to certain scientific traditions or even individual predilections). In the edited volumes this random kind of variation has been strained off and the user is presented with the significant forms and variation. To this should be added that a majority of the paper slips are hand-written and often quite difficult to read – in general the information on the paper slips can be rather sketchy due to the practical circumstances when the information was taken down in the field.

One of the strengths of digitization is that it makes it possible to make inside and/or densified information more accessible and explicit. However, given the limited resources available we are considering to what extent we should make a raw publication a priority, exactly because a lot of the information on the paper slips would require an extensive work with specialized, often internal metadata and incomplete standard data. And perhaps more important: are these kinds of information relevant or helpful for possible future users? A number of these users will have less specialized qualifications but they will often have a keen and vivid interest in the key information dealt with in the DID project. It will often be difficult for these users to find the interesting information among the sometimes numerous information types on a single paper slip presented in isolation, and they will not get more generalized information about the significance or contextual value of a given information on a single paper slip – the sought-after information would be blurred and difficult to grasp. Generalized and salient information, however, can be found in the edited data – and incomplete standard data on the paper slips will also have been dealt with here.

5 Conclusion

In a future work with metadata and use/publication potentials we anticipate a division between salient key metadata and more specialized metadata. Salient metadata are essential to finding and understanding the main content in the collections and play a key role in giving different users, internal as well as external, specialized as well as less specialized or common or garden, access to the sub-collections. These data are also important in order to link the sub-collections together in a dynamic and meaningful information architecture. Specialized metadata are important for more specialized uses but less important, sometimes perhaps even confusing and counterproductive, for less specialized users and less specialized search purposes. It might be relevant or worthwhile, especially in a priority situation with limited resources, not to present certain types of digitized data for a wider public in an un-edited form. Consequently, the edited volumes, especially the metadata in the xml schema in the new volumes, play a key role in the future work with the collections and the dissemination of information in the DID project.

Bibliography

Bakken K, Grønvik O (2008) Materialsortering på digital platform. Eit steg mot høgare dataintegritet i den vitskaplege leksikografien. In: Svavarsdóttir A et al. (eds) Nordiske Studier i Leksikografi 9. Skrifter udgivet af Nordisk Forening for Leksikografi 10. Reykjavík, pp 31–42.

Engerer V, Roued-Cunliffe H, Albretsen J, Hasle P (2017) The Prior-project: From Archive Boxes to a Research Community. In: Digital Humanities in the Nordic Countries, 2nd Conference, Göteburg, March 14–16 2017. Gothenburg, pp 53–57.
Grønvik O (2016) Vitskaplegheit og samfunnsrelevans for store ordboksverk. In Gudiksen A, Hovmark H (eds) Nordiske Studier i Leksikografi 13. Skrifter udgivet af Nordisk Forening for Leksikografi 14. København, pp 27–61.
Tasovac T, Petrović S (2015) Multiple Access Paths for Digital Collections of Lexicographic Paper Slips. In: Kosem I et al. (eds) Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom. Ljubljana/Brighton, pp 384–396.