LIBER QUARTERLY, ISSN 1435-5205

K.G. Saur, Munich. Printed in Germany

Digitising Complex Data: Integrating Text, Images and Physical Organisation

by ESPEN S. ORE

BACKGROUND

The National Library of Norway (NLN) is the Norwegian national deposit library and has a history going back to 1815 when national library functions were taken care of by The University of Oslo Library. In 1989 the Rana Division of the National Library was established and given the responsibility for the administration of the Norwegian arrangement of legal deposit. In 1992 the Norwegian Parliament decided to establish a unified self-contained National Library with its main office located in Oslo. The Office of the National Librarian came into operation in 1994. The Oslo Division (NLN-O) was established in 1999 based on national functions and collections separated from the University of Oslo Library.

SOME COLLECTIONS AT THE NLN-O

The Manuscripts collection include manuscripts, letters, and archival material from Mediaeval time to the present with main emphasis on 18th and 19th century materials from Norwegian scholars and men of letters. The Picture Collection includes about 500,000 objects, mainly portraits and topographical photos. The Map Collection includes about 150,000 objects, comprising all maps printed in Norway, manuscript maps, foreign maps, and printed material on cartography and topography. The Drama Collection includes archival material from Norwegian theatres and printed material focusing on the dramatic arts. There are special collections covering emigration history and Norwegian settlements in North America, the Norwegian participation in the Second World War, and the Norwegian authors Henrik Ibsen and Bjørnstjerne Bjørnson. The NLN-O also holds the music collections including printed music (about 200,000 units), manuscript music (18,000 units) by Norwegian composers, sound recordings of Norwegian music (37,000 units), archival material on Norwegian music, books and serials related to Norwegian and foreign music.7

DATABASES

The NLN are currently providing internal and external access to more than one hundred separate databases. Some of these databases have a history at least thirty years old, and in some cases the software used goes back to the same time. In the following I will describe and discuss problems related to special collections databases at the NLN-O. For many reasons such databases have been developed as freestanding tools for the separate departments and for the different types of collections. In some cases the same types of information such as names or places are used in different collection databases. In some cases also the same types of objects (for instance posters or pictures) belong to different collections and have up until now been recorded in different databases - as far as they are registered in databases at all.

At the NLN-O we have started work on a strategy for unifying the separate databases. However, since the collections to a large extent are not catalogued in a database at all we are also making pilot databases where we use selected data from the different collections.

VIRTUAL COLLECTIONS

Apart from the databases that catalogue the objects in our collections we have also started to digitise selected objects. A collection of for instance photographs may seem fairly straightforward to digitise: one digital copy made from each existing photograph or negative. This gets more complicated if we also consider different digitisations of the same object as objects in their own right but I will not go further into that. For some objects we have the additional complication that they are in themselves collections of objects. This has consequences both when it comes to cataloguing and when it comes to making digital copies available: how is a user interface meaningful with such complex objects? When we are making digitised objects available, should we also make text available as text? Should we prepare transcriptions, and in that case should we select an encoding standard? One item in the manuscript collection illustrates some of work we have done so far:

THE DIRIKS FAMILY SCRAPBOOKS

The Norwegian Diriks family scrapbooks, a part of the manuscript collections at the NLN-O, comprise 13 large bound volumes (albums) and three folders of loose material covering a time span of more than a 100 years from the be-

ginning of the 19th century to the beginning of the 20th century. The material was collected and organized by Anna Diriks (1870-1932). The contents of the scrapbooks vary and include letters and notes, newspaper clippings, drawings (including original drawings by artists such as Max Klinger and Amedeo Modigliani.

The Diriks scrapbook albums need to be conserved. The albums are of cheap quality from around 1890 and the pages have a high acid content. The material in the albums is much sought after by researchers in history, art history etc. and for use as illustrations in publications. For these reasons the NLN has decided to:

- do a physical conservation of the scrapbooks,
- register the contents (objects) of the scrapbooks in a database,
- digitise the scrapbook contents,
- produce a virtual copy of the original albums on the web.

A pilot project was done in 2000 and a production project based on the experiences from that pilot project ended in March 2002 resulting in one album (1a) conserved, digitised and partly published on the web (DIRIKS-2002). The remaining 12 albums will be conserved and published over the next 6 years depending on the available funding.

This project is inter-departmental at the NLN-O: the Manuscript department is responsible for registering the contents (objects) of the volumes; the Conservation department does the conservation work and in collaboration with the IT department administers the digitisation of the pages and their objects. The IT department is responsible for the database development and for the web version of the scrapbooks.

The publication of the scrapbooks on the web requires manual adjustment of the digitised objects and allows us to automate only some tasks. The object-browsing model (see below) however allows us to place the edited digital objects in the data structure more or less automatically. The database on the other hand automatically serves as a basis for many purposes:

- It documents all objects.
- It links conservation information with the objects.
- It links digital images with the individual objects.
- It is available for the web version, and so allows for searching and other kinds of retrieval than virtual browsing through album pages.

What Objects are there in the Diriks Scrapbooks?

On a given album page there may be text and drawings directly on the page (see Fig. 1). In addition there usually is one or more items adhered to the page. Such an item may itself be of more than one page (for instance a printed or handwritten booklet, a multi page letter or a set of newspaper clippings) and these items and pages may have other items adhered to them.

In the pilot project in 2000 a data model of the albums was set up showing a hierarchy:

- Scrapbook album
- Album pages
- Objects on album pages

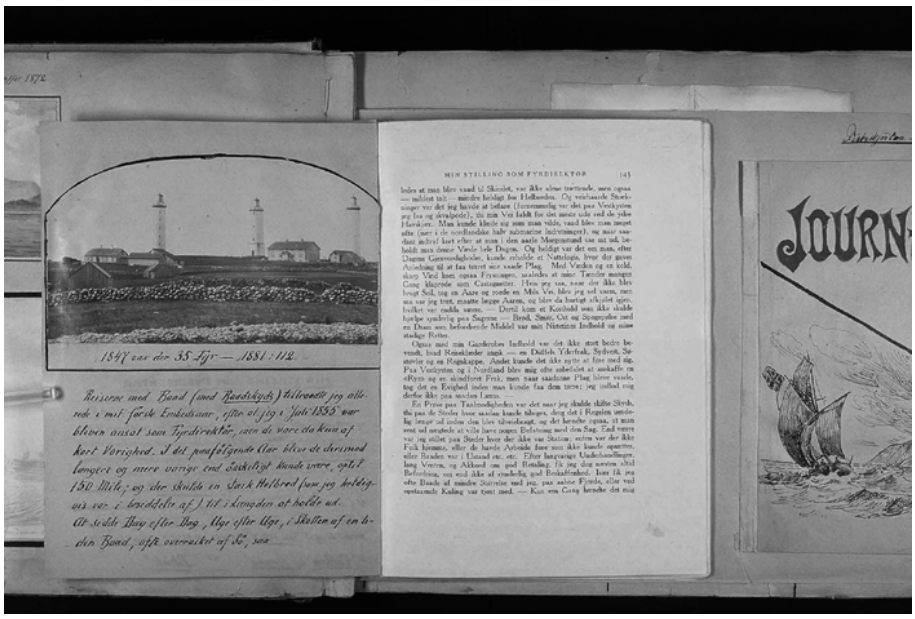


Figure 1: from Diriks album 1a, 6 recto

The work done in 2001 has shown that this model was too simple or maybe too confining. The 2000 work did not include conservation information in the database model, and the example volume (album 5) did not have drawings and text directly on the album pages in such a way that it was felt necessary to

keep them intact or as part of the data referenced from the database, except for on an abstract top level. The database was in 2001 extended to allow for conservation information. The data model was changed to allow us to treat albums and album pages as objects in the same way as the objects adhered to the pages. This made it necessary to introduce a notion of „is part of" and „holds the following objects" - something which in any case would have been needed also for the original and simpler data model once registration of objects further down in the hierarchy than at the page level started.

Conservation, Preservation and Digitisation

Before the conservation work the albums and their pages undergo a first level of photographing/digitisation. These images are linked with the album objects in the database and are used as a tool during conservation. The images also document the original album pages and are used as a basis for building the web-version, a virtual preservation of the albums. When the objects on the pages (and the pages themselves) have been through the conservation process they are digitised on an individual level. These images are also linked with the database and in various resolutions they are used as browsing objects in the web version of the albums. The high-resolution images are also intended for researchers and for publication purposes. Ideally they can be used as replacements for the originals thus reducing the handling of them and so hopefully aid the preservation of the conserved objects.

The Web Version

For the web presentation of the scrapbooks a restricted view of the database is available. This of course includes the registered names, places and dates as well as the listed object types. There are also additional introductory texts. Gateways to the data in the form of a time line and possibly a map are considered. Ideally the text in or on the objects should be available as encoded full text and it should be searchable. This has been set aside due to lack of funding. Some of the texts and person names are extracted and stored in the catalogue database, however.

One of the important aims for the project however is to give a digital presentation of the albums such as they were before treatment since the organization itself gives information about the objects and their relationships. The albums and the organization of the collection are also objects of cultural historical value by themselves.

Since a given page may hold one or more objects, which may in themselves be multi page and hold other objects it is necessary that the reader should be able to browse in two dimensions. We have designed this browsing model:

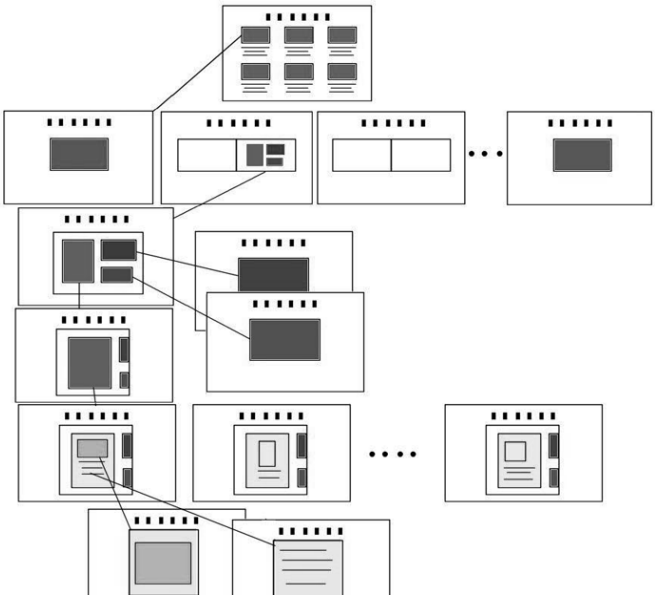


Figure 2: The model for browsing a virtual Diriks scrapbook (designed by Jingru Høivik at the NTN-O)

This model is realized thus:

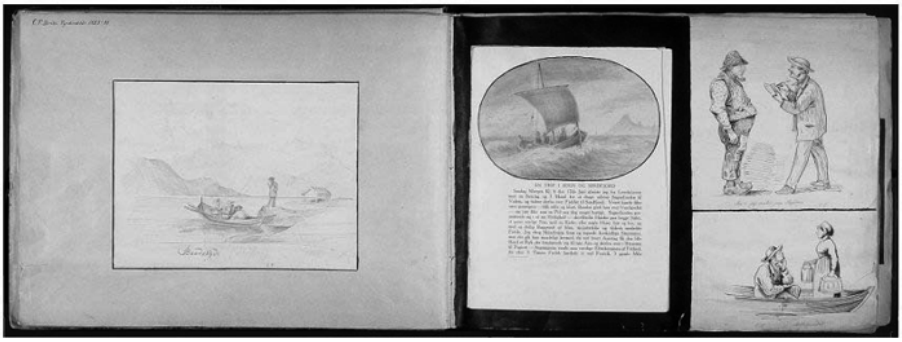


Figure 3: Pages 4 verso and 5 recto, Diriks scrapbook 1a, a view of the opened album showing two opposite pages.

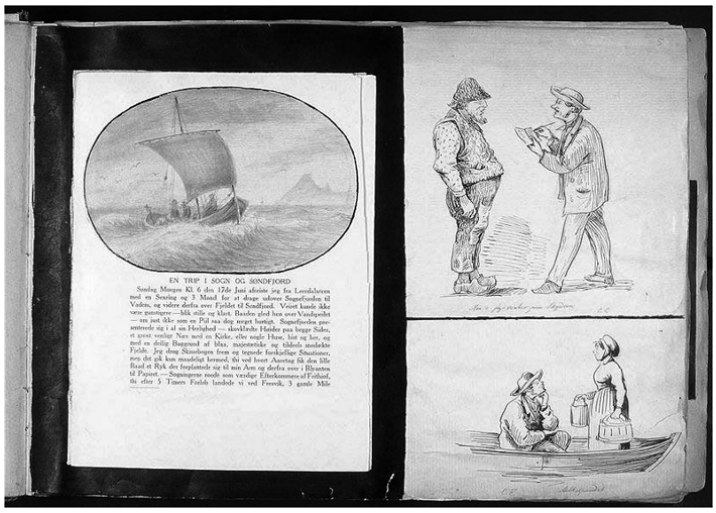
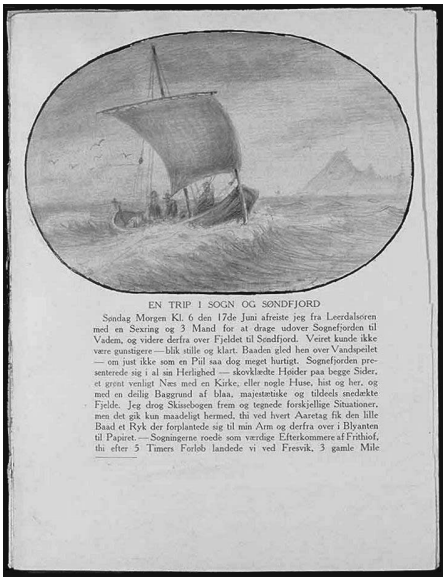


Figure 4: Page 5 recto, Diriks scrapbook 1a. Here one page has been selected.



EN TRIP I SOGN OG SØNDFJORD

Søndag Morgen Kl. 6 den 17de Juni afreiste jeg fra Leerdalsören med en Sexring og 3 Mand for at drage udover Sognefjorden til Vadem, og videre derfra over Fjeldet til Söndfjord. Veiret kunde ikke være gunstigere — blik stille og klart. Baaden gled hen over Vandspeilet — om just ikke som en Pil saa dog meget hurtigt. Sognefjorden præsenterede sig i al sin Herlighed — skovklædte Heider paa begge Sider, et grønt venligt Næs med en Kirke, eller nogle Huse, hist og her, og med en deilig Baggrund af blaa, majestætiske og tildeels snedækte Fjelde. Jeg drog Skissebogen frem og tegnede forskjellige Situationer, men det gik kun maadeligt hermed, thi ved hvert Aaretag fik den lille Baad et Ryk der forplantede sig til min Arm og derfra over i Blyanten til Papiret. — Sogningerne roede som værdige Efterkommere af Frithiof, thi efter 5 Timers Forlob landede vi ved Fresvik, 3 gamle Mile

Figure 5: Detail from page 5 recto, Diriks scrapbook 1a. One object on the page has been selected. Notice, however, that this object in itself comprises many pages and that the drawing above the printed text on this page is a separate original drawing which has been pasted onto the printed page.

HOW TO BROWSE DIGITISED MATERIAL?

Virtual Albums

In the Diriks project we have chosen a „turning the pages" metaphor for the browsing interface. One of our principles is that as far as possible our data should be available for the public using standard software (i.e. web browsers). This does not mean that we will not publish CD-ROMs with software included (such as the British Library's Electronic Beowulf edition (BW-2002) or the OUP/University of Bergen edition of Wittgenstein's Nachlass (WN-2002)) but that kind of products should be a spin-off from our general work in producing catalogues and digital facsimiles.

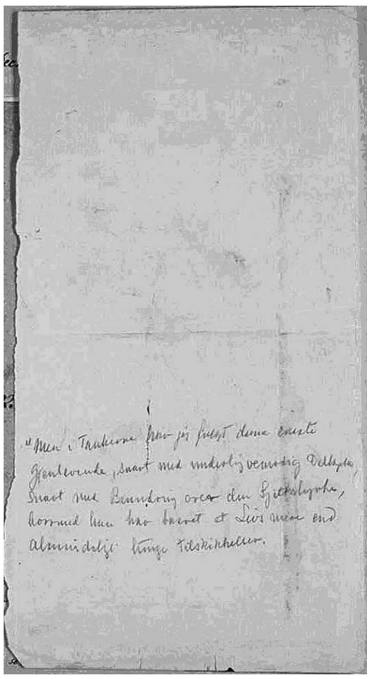


Fig. 6

The model of turning pages is not without problems. How do we model a scroll? In the Diriks scrapbooks there are objects with text and/or drawings on the backside. When such an object has been pasted onto an album page or another object the backside has usually not been available to a person browsing the album – we are providing something that was not there in the virtual version. In the model we have set up now there is a link informing the user that there is something on the backside.

Full Text and Catalogues

To use the British Library once more as an example I will point to the International Dunhuang Project (DH-2002). Here access to an individual object is through catalogue information or finding aids. This database is obviously not intended for the general public, but still the user interface with thumbnails showing the digitised pages available and with facilities for zooming. This works in practice like a combination of a catalogue, a database and the collection of objects and is probably very good for scholars who know what they are looking for.

A similar approach to the one in the Dunhuang project has been used by the University of Oslo and the NLN-O in a collaboration project where all known (and available) manuscripts and letters in Henrik Ibsen's hand were digitised (HIMS-2002). Here the overall access is based on work (play/poem) or date (letters). Encoded transcriptions are being made available (in a separate project) and will at some time in the future be linked with the facsimiles in a searchable version (HIW-2002). An early version of the transcribed texts are already available (HIT-2002) as part of a 60,000 page Norwegian literature corpus handed over to the NLN from the University of Oslo (DOKPRO-2002).

A third example can be found at the Arnamagnæan Institution in Iceland (Stofnun Árna Magnússonar, AM-2002). Here we can find selections of digitised Norse manuscripts. And again the access is through a catalogue rather than a text search. (See also Driscoll 98)

If I should dare to look into the future I would say that I see a development towards full text searching tools. Making a manuscript collection available for free text search makes it available in a way that is qualitatively different from access through a catalogue or finding aids in a database.

DIGITISING OBJECTS

When we are digitising the collections at the NLN there are certain decisions which have to be made: are we digitising for eternity or for a publication? Do we worry about how the digital copies may be used? What purpose are we publishing them for - if we are? These considerations can be listed up roughly as:

- Resolution and image quality for public web presentation
- Resolution and image quality for e.g. high quality print

- Resolution and image quality for archival functions
- Resolution and image quality vs. copyright and related matters

The electronic publication of Wittgenstein's Nachlass has shown that his handwriting is more than acceptable for reading and scholarly work if a digitised version is presented in twice real size (in practice a resolution of 150 dpi when the material is scanned) (Ore&Cripps, 98). This resolution is also acceptable for simple print purposes but not for high quality facsimile prints. But recent experiments at the NLN-O have shown that writers from other Western-European traditions such as French early 19th Century hands may require a larger resolution. On the other hand some web sites present digitised material with a lower resolution than 150 dpi, and then the material quickly becomes unsuitable for anything but purely illustrative purposes.

Photographs have varying and individual requirements when it comes to resolution and image quality but again a twice real size from a print will usually be more than good enough for web presentation.

Objects with line art, including maps, have their own problems. At times it will be preferable to enhance lines on a digitised map - the alternative is storing the image with a very high resolution.

IMAGE DATABASES

The NLN is now in a planning phase when it comes to image databases. A simple generally available image database has been in use for some years now (GN-2002) but the internal database from which this public version has been extracted has so far been a temporary ad hoc solution.

UNIFYING DATABASES

For almost a year the NLN-O has used a manuscript catalogue database (HANSKE-2002). The Diriks scrapbooks described above would as a matter of fact be registered here as one single object. However, the separate image objects found on the album pages should also be registered in an image database. This shows us the need for a top-level data model independent of departments and collections. By integrating what were once separate independent databases it will also be easier for us to provide our users with general, cross-collection searching tools.

REFERENCES

(All URLs given here were checked on June 20, 2002)

AM-2002, <http://am.hi.is/skrift/test/valmynd.pl?>.

Burnard, Lou, Marilyn Deegan and Harold Short (eds): *The Digital Demotic,* A selection of papers from the Digital Resources in the Humanities 1997, Office for Humanities Communication, King's College London, 1998.

BW-2002, <http://www.bl.uk/collections/treasures/beowulf.html>.

DH-2002, <http://idp.bl.uk/>.

DIRIKS-2002, <http://www.nb.no:9000/utv_nbo/jh5/index.html>.

DOKPRO-2002, <http://www.dokpro.uio.no/litteratur/>.

Driscoll, Matthew James: „The virtual reunification of the Arnamagnæan Manuscript Collection", in *The Digital Demotic,* 1998.

GN-2002, <http://www.nb.no/gallerinor/>.

HANSKE-2002, <http://www.nb.no/hanske/>.

HIMS-2002, <http://www.dokpro.uio.no/litteratur/ibsen/ms/indexe.html>.

HIT-2002, <http://www.dokpro.uio.no/litteratur/ibsen/>.

HIW-2002, <http://www.ibsen.uio.no/his/hjemmeside/english.html>.

Ore, Espen S. & Peter Cripps: „The Electronic Publication of Wittgenstein's Nachlass" in *The Digital Demotic,* pp. 111-118, 1998.

WN-2002, <http://www.oup.co.uk/academic/humanities/philosophy/wittgenstein/>.