Knowledge Exchange

A snapshot of digitisation and OCR developments

Survey and synthesis performed in 2014

DFG Jisc CSC DEff SURF

1. A KE questionnaire on digitisation?

One of the goals of Knowledge Exchange (www.knowledge-exchange.info) is to help partners to share knowledge and expertise and facilitate the build of expert networks. In the area of digitisation, including non-textual and 3-D-digitisation, a first step is to provide a snapshot of current activities and challenges in the KE partner countries.

This paper is a synthesis of the information gathered in a questionnaire that was sent to 15 infrastructure institutions, e.g. libraries and also funders, within the five partner countries of Knowledge Exchange: Denmark (DK), Finland (FIN), Germany (GER), the Netherlands (NL) and the United Kingdom (UK).

The paper is based on the answers provided by 6 respondents from four countries:

- **DK**
  - Danish Agency of Culture (Henrik Jarl Hansen)
  - State and University Library (Tonny Skovgård Jensen)
- **GER**
  - German research foundation, DFG (Franziska Regner)
- **NL**
  - Royal Library (Hildelies Balk)
  - Leiden University Library (Saskia van Bergen)
- **UK**
  - Jisc (Paola Marchionni, Peter Findlay)

The absence of Finnish responses may be due to Finland participating in the recent Enumerate Core Survey II that also addressed digitisation. We have included some of the outcomes of this survey to present a richer picture.

Informative and indicative

The emerging picture is primarily meant to inform the partner organisations. They may find outcomes that are of mutual interest and may get in touch with the identified initiatives in the digitisation area. KE gladly shares with those who are interested in the area.

The results are by no means sufficient to represent the actual situation in the five countries; the synthesis identifies the sort of activities and challenges currently at play. Similarities and differences noted, are based on a limited number of respondents, and are indicative only.

Following the production of a preliminary report, respondents were given an opportunity to provide further clarifications and to elaborate on their initial responses. This additional information has been added.

2. Synopsis of responses to questions on digitisation

The number of responses to the twelve questions does not allow for quantitative analysis. However, we tried to synthesise the answers to provide a more telling picture.

Question 1: Which types of institutions digitise texts and objects in your country?

In all four partner countries three types of institutions are mentioned most often: firstly the national and university libraries; secondly archives; and thirdly museums.

Denmark and the UK mention other actors as well: various archives (regional, local, specialised), public libraries, conservatories, theatres, academic departments, publishers (also newspapers, music and broadcasters), private collectors and companies, sometimes in public private partnership.

These outcomes correspond with the Enumerate survey – digitisation in Finland happens in a variety of libraries, museums, archives and other institutions.

Question 2: Who finances these digitisation activities mostly?

Three main sources of financing for digitisation can be found in all four countries. First, there are government funds, either directly (e.g. projects) or indirectly provided by councils and ministries. Secondly, there are research funding bodies, either public (e.g. DFG) or private (e.g. Wellcome Trust). Thirdly, there are commercial activities, on the one hand big players (e.g. Google, BBC, Proquest, Cengage, Brightsolid or publishers in general) and on the other hand public private partnerships (mostly with Google). In Denmark for example, there is a huge interest, both on the collection side and on the user side (research and education), to use digital collections of TV and radio broadcasts.

Digitisation is also often financed by institutions through their regular budget and on a much smaller scale, by private sponsors, federal states, regional and local public funds, small funding organisations (e.g. Volkswagen Stiftung), private companies and by customers for digitisation on demand.

The Enumarate survey reveals that funding for digitisation in Finland mainly comes from internal budgets and national public grants but other (e.g. regional/local, public-private, commercial) sources as well.

Question 3: Is there a large interest in digitisation? And how is that for non-textual materials?

The answers from all four countries to this question are all similar to a point; there is a large interest in digitisation of both textual and non-textual materials however if one looks at the answers in detail, there are differences in terms of who is identified as having an interest in digitisation.

We could differentiate between those who are interested in the process of digitisation and those who are interested in use and access of digitised material. The interest in the process of digitisation is referred mostly to librarians, specialists and scientists who are working on the technology for digitising. With regard to the use and access of digitised material, the responses reveal at least two types of interest. First the user-interest (i.e. mostly scholars and scientists, journalists and the greater public) and second the political interest (an increased awareness can be observed). Notably in Germany the significant interest in non-textual objects for scientific collections comes from both policy level as well as the research community, eg. the 3D scanning system CultLab3D of Fraunhofer IGD that will be tested at the Museum of Natural History in Berlin (for more information, follow this link).

Furthermore, there is a wide variety to the type of the non-textual materials. For example in the UK the interest in digitisation applies to audio, film and TV materials. This also applies to Denmark that also has an interest in digitisation of maps, museum artefacts and documental material (i.e. from archaeological excavations).

Question 4: What challenges are you facing regarding the automated retrieval, quantitative analysis through text or data mining, semantic analysis, pattern recognition in non-textual materials, data enrichment, contextualisation and further processing of digitised material?

The main challenges identified regarding the overall process described in the question are legal hindrances. Copyrighted materials require complex licenses to make them available for the listed techniques. This holds true for all four countries.

For each country, additional obstacles apply. The answers for Denmark suggest that the museum sector has not yet reached the stage that the question refers to. Within the library sector however, a major challenge is the (un)availability of sufficiently homogeneous metadata for search, retrieval and analysis purpose. Answers from the Netherlands refer to a project on text mining that just started. A bigger challenge within textual material is the poor quality of OCR and pattern recognition in historic material. For Germany a main challenge is licensing of full-texts or image-related data for full reuse with the above techniques. Another challenge is the improvement of pattern recognition in non-textual materials. For the UK, Jisc points at the immense value of allowing large data sets to be data mined but that most projects have encountered rights issues. Additional barriers to engaging with digitised content in richer ways are: the unsatisfactory level of discoverability of much digitised content (see the findings and conclusions of the Jisc Spotlight on the Digital project, the general fragmentation (silos) of datasets and collections especially within the humanities; the underlying quality of data and metadata; the lack of appropriate digital literacy skills to enable meaningful interaction with digital content within staff (Higher Education teachers, researches and academic support staff such as liaison librarians) and students.

Question 5: What standards are there to catalogue and digitise non-textual material in scientific and museum collections?

In general, no common standards for digitising non-textual material can be identified, neither within a country nor within the international answers to this question. The specifics for each country are listed below.

DK	• local standards • establishing at the moment international standards (CIDOC, UK Spectrum, Europeana Data Model) • PB core, metadata specification of Europeana Sounds
NL	• MARC21 • METS • TEI • EAD • JP2 • CCO • CDWA lite • JPG, TIF
GER	• no common standards for covering all non-textual materials • LIDO core elements • DFG practical guidelines on digitisation
UK	• A variety of standards are used in the UK, both sector-and format-specific. As an example, please see the guide from Jisc Digital Media on Metadata Standards and Interoperability

Question 6: Are there legal hindrances to be tackled to provide researchers with fully reusable digitised material – especially in the European and international context? What is to be done to overcome these hindrances – both on the political level and on the practical level?

The copyright issue is mentioned in all answers. Within a single country there are legal frameworks to legalise the use, but providing digitised material within the European and international context is much more difficult. On a political level all the respondents wish for more activity undertaken by the European Commission and international bodies. On a practical level the answers propose to use open access wherever applicable, promoting unrestricted use and access as well as the use of Creative Commons Licenses.

Question 7: Would you consider the collection-level-description of digitised collections state-of-the-art in your country? And why?

The answers indicate that there are two distinctive groups; on the one hand the Netherlands and the UK with more or less well discoverable nationwide high-level metadata. The results of

the Jisc Spotlight on the digital project revealed that on the whole there is good discoverability of digitised collections at collection level. However, there is a serious problem of discoverability at item level, whereby a majority of digitised items are not surfacing through search engines searches (for more information see this blog post) On the other hand answers from Denmark and Germany refer to a lack of machine-readable and standard-oriented data on the collection level in repositories.

Question 8: How is long-term-archiving of digitised materials taken care of?

The long term-archiving is done either by an institution itself, by a national library or by a specialised data center. Information, practical guidelines, standards, recommendations and policies are provided by funders or national libraries.

However, there are several issues regarding different kinds of digitised materials, e.g. for Denmark geographically related materials. For Germany there is currently no universal solution that is suitable for all types of objects and materials. Another challenge, mentioned for the UK, is how to ensure that institutional repositories are interoperable.

For Finland the Enumerate survey shows that a small part of the Finnish digital collections is considered to be in an archive that meets the international criteria for long term preservation, and a significant part is archived in a publicly or privately professionally managed digital archive. Most Finnish respondents say they have no solution for long-term archiving.

Question 9: Are you of the opinion that there are good solutions for long-term-archiving of digitised materials in your country?

The solutions are developed by different actors, they are

NL: national depot, or the use of use DANS http://www.dans.knaw.nl/ and http://www.3tu.nl/en/

UK: the developments of DPC and DCC

DK: The Digital Preservation Strategy is rated as very good, but it is a specific, not a national strategy.

GER: According to DFG's practical guidelines on digitisation the Open Archival Information System (OAIS) should be used as a reference model for the archiving of electronic data. The 'Criteria for Trusted Digital Repositories' are essential.

It should be noted that different types of digital preservation strategies are mixed here: strategies for national digital collections of cultural heritage like newspapers, radio and TV broadcasts, and preservation strategies aimed at research data.

Question 10: What solution would you consider best?

The respondents were not able to answer this question or mentioned their answer to the last question.

For Germany: There is no "one size fits all"-solution. Commercial products, such as Rosetta by ExLibris, are being implemented in some larger institutions. They might provide stable solutions.

3. Synopsis of responses to questions on OCR

Question 11:

Best practices in Optical Character Recognition (OCR)-projects: Are there competence centres dealing with the OCR of prints of specific centuries or printed in specific fonts?

All respondents referred either to commercial vendors or to IMPACT Centre of competence in digitisation. In Denmark and Germany there are also in-house competences. DFG for example initiates activities to further improve OCR practice.

Question 12:

What are best practices to define OCR accuracy?
Are there established standards in your country?

On the one hand the respondents referred again to IMPACT Centre of competence in digitisation. On the other hand it is discussed whether this is a question of defining or measuring. The answer from Denmark explained that they did not find any established and usable standard. They said it is rather hard to evaluate the accuracy except by sampling. The situation is very similar for the UK, where it is argued that accuracy depends on the input material. The best practice from the perspective of the Netherlands is to look at word accuracy rather than character accuracy, and to do proper evaluation with ground truth instead of relying on the "assumption" of correctness of an OCR engine. For Germany the DFG practical guidelines on digitisation provide best practice advice: To check the accuracy of transcribed or OCR-generated texts, statistical methods must be applied. The aim is to assess, on the basis of a random sample, whether the recognition rate claimed by a service provider can be relied upon. The probability of error should be kept as low as possible while keeping the size of the random sample manageable. The statistical method required is a so-called "Bernoulli trial". Another DFG initiative '*Weiterentwicklung von OCR-Verfahren'* aims to provide further guidelines for OCR practice.

Sources

The Enumerate Survey II report can be found here.