THE DIGITAL FACTS OF CULTURAL HERITAGE

Marco de Niet

The DEN foundation, P.O. Box 90407, 2509 LK The Hague,
The Netherlands - marco.deniet@den.nl

KEY WORDS: Digital Heritage Statistics; NUMERIC; SIG-STATS; ENUMERATE; Netherlands; Monitoring; The Digital Facts

ABSTRACT:

This short paper presents some results of research into the current practices of digitisation by cultural heritage institutions across Europe. The paper addresses activities that were set up in the context of the EU-funded project Numeric (2007-2009), such as the Special Interest Group on Cultural Heritage Digitisation Statistics. The paper will focus in particular on monitoring activities conducted in the Netherlands. The Netherlands were able to provide reliable data from 131 heritage institutions to Numeric, which turned out to be the largest contribution of all participating countries. These monitoring activities went beyond the scope of Numeric and also included topics such as born digital heritage that were not addressed in Numeric. Finally, a possible follow up to Numeric will be discussed.

1. INTRODUCTION

In recent years, there has been an increase in the need for intelligence about the progress of digitisation of cultural heritage. Since the early nineties, most cultural heritage institutions have engaged themselves in digitisation projects, ranging from mass digitisation projects of newspapers and audio-visual materials to small scale activities, for example to promote the highlights in a specific collection. Many of these projects were financed with public money, often through additional funding from local or national governments, public funds or the European Commission. What can we say about the total amount of cultural heritage that has been made available digitally, online or on site? How much has been invested so far? And is there any reliable data to be given about the use of the digital collections? If we are able to provide answers to questions like these, we will be able to better plan and manage our future digitisation activities.

Monitoring digitisation activities has been done for quite some time now. In Europe, many European FP7, eContentPlus or ICT-PSP projects start with a survey to determine the current status or size of digital collections, the use of metadata schemes and other standards, or the availability of databases and other services. The Lund meeting from 2001 in particular led to a better common approach to quality assurance for the use of ICT by cultural heritage institutions. In various EU-countries surveys were set up to understand better the trends and needs of the institutions and their users. International initiatives like EGMUS (European Group on Museum Statistics) included data on ICT and digitisation as well. However, an overall methodology for monitoring the progress of digitisation of cultural heritage was lacking. It was considered useful for both policy makers and institutions to set up a large scale survey that would define the empirical measures for digitisation activities and establish the current investment in digitisation and the progress being made by Europe's cultural institutions. This became the Numeric project.

2. NUMERIC

Between 2007 and 2009, the European Commission contracted UK-based CIPFA (formerly The Institute for Public Finance) to undertake the Numeric study to

1. test a framework for collecting and analysing data relating to digitisation activities of materials held by libraries, archives and museums in the EU and
2. implement this with the help of nominated experts in each European Country.

Numeric was a groundbreaking effort to collect and harmonise statistical data on digitised cultural heritage across all EU-member states. The key instrument was a rather extensive questionnaire, which addressed topics like information policies, size of collections, investments, staff involvement, use of standards and usage of digital collections. To support this survey tool, Numeric developed other instruments, such as a terminology list and a tool to determine a representative sample of cultural heritage institutions in each country.

The two most important results of the Numeric project were the Study Report (published as a draft in May 2009, the final version was published in February 2010), and the Numeric Framework, a group of institutions and persons, brought together through a shared interest in the Numeric objectives. In each EU-country, a National Coordinator was appointed, usually by the Ministry of Culture. These National Coordinators were instrumental in promoting the Numeric Survey across Europe and involving heritage institutions to contribute their data.

In total, 788 respondents from 26 countries participated in the Numeric Survey. In itself quite a respectable number, which provides a proper foundation to the facts and figures presented in the Study Report. Here are two interesting outcomes of the Numeric Study:

Table 15 Progress made towards the digitisation of collections

Type of institution:	Part of collection digitised % [1]	'Order book' Completed % [2]	'Order book' Outstanding % [3]	Equivalent backlog [3]/[2] [4]
Archives	5.1	10.3	89.7	8.7
A-V or film institutes	9.8	15.4	84.6	5.5
Broadcasting institutes	10.8	12.8	87.2	6.8
Art/archæo museums	27.2	30.6	69.4	2.3
Science and tech museums	25.5	32.4	67.6	2.1
Other museums	17.5	23.1	76.9	3.3
National libraries	2.3	3.5	96.5	27.6
Higher education libraries	2.5	4.4	95.6	21.9
Public libraries	14.8	31.9	68.1	2.1
Special or other libraries	5.5	12.2	87.8	7.2
Other types of organisation	22.5	29.0	71.0	2.4

Figure 1: Table 15 from the Numeric Final Report

Table 15 provides an overview of the average part of the collection that has been digitised so far. The 'Order book' refers to the part of the collection that an institution intends to digitise. With the equivalent backlog, it can be calculated how much more cultural heritage needs to be digitised. For example: The amount of archival collections to be digitised equals the current size multiplied with a factor of 8.7.

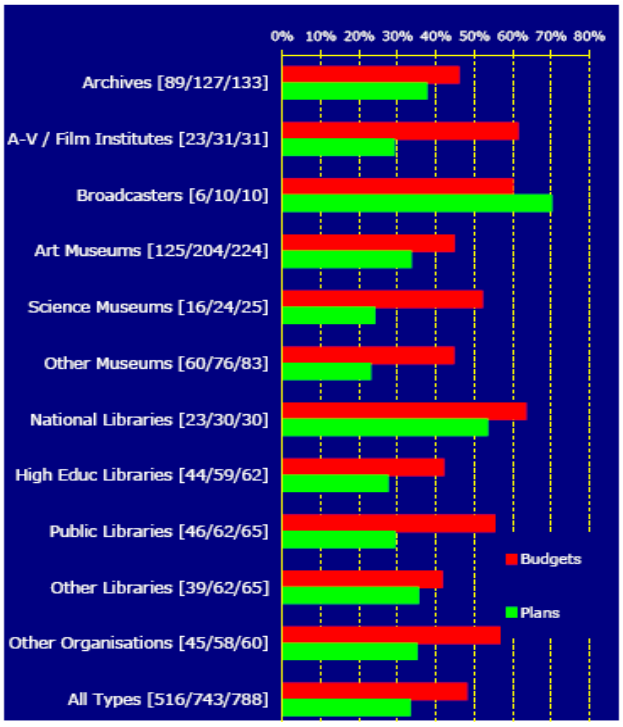

Figure 2: Figure 5 from the Numeric Final Report: "The first two figures in brackets indicate the number of institutions respectively responding to the question about their (1) BUDGET and their possession of a (2) PLAN; the third figure is the total number of (3) survey responders. Some will not have indicated that they possess a budget or a plan; the proportion that did is indicated by the bars in the chart."

This second graph shows how many institutions have a policy plan on digitisation in relation to the amount of institutions that have a specified budget for digitisation. With the exception of broadcasters, all cultural heritage domains have (considerably) more specified budget than they have policy plans. In short: there is a lot of 'ad hoc' digitisation going on.

Relevant and interesting as these results may be, it has to be said that the responses from across the EU to Numeric were rather unbalanced. With 131 contributing institutions, the Netherlands provided the largest set of data, while countries with much larger cultural heritage communities provided far less data (from Italy and France, only 30 institutions participated in Numeric). From the beginning, Numeric identified commitment from heritage institutions as a key factor to success. During the project, two workshops were organised for the National Coordinators, to discuss the methodology, approach, and the preliminary results and future options. One of the recommendations to emerge from these workshops was to set up a Special Interest Group, dedicated to developing the standards and definitions for future survey activities.

3. SIG-STATS: recommendations for a follow up

The National Coordinators from six EU-countries volunteered to set up this Special Interest Group, called SIG-STATS: Austria, Belgium, France, Germany, Hungary and the Netherlands. The installation of SIG-STATS was endorsed at the meeting of the Member States Expert Group, on 1 October 2009. In February 2010 the SIG met in Luxembourg to prepare the recommendations for a follow-up to the Numeric study. The SIG addressed seven topics that needed closer attention:

1. Survey design: principles to structure the questionnaire(s)
2. Defining the survey sample: the criteria for identifying 'relevant' institutions
3. Definitions: improving the 'vocabulary' of the questionnaire and harmonising it in the EU27 languages
4. Input-output measures: how to link the analogue heritage to the digital representations
5. Calculation of costs
6. Measuring usage and access: valid and practical means of measuring access to and use of digitised heritage
7. The framework: organisation and implementation of the survey across the EU27 countries

For each of these topics, the SIG discussed possible improvements. A key principle for the SIG was that the follow-up should not only be about the gathering of data about the here and now; it should also support the heritage institutions to get more 'in control' of their digitisation activities, by showing them the usefulness of having better intelligence about the size, costs and usage of their digital collections. The follow-up to Numeric should not be only about short term statistics, but also about long term accountability and performance indicators. The consequence of this principle was that it should be accepted by all parties involved that it will take a few more years of research before we have useful benchmarking data.

In short, the SIG recommended a hybrid approach to surveying current digitisation practices, which on the one hand will not compromise the original goals of the Numeric survey (i.e. to get a better understanding - from a policy point of view - of the growth of and investments in digital cultural heritage) and on the other hand will appeal to the institutions to participate in their own interest. The three main topics of the Numeric survey (size, costs and usage of digital heritage) are the results of complex sets of activities and procedures, and the SIG assumed that most of the institutions are still trying to make these activities run smoother and more efficiently. By analysing the 'digital workflows' in the cultural institutions, not only more precise definitions can be obtained for surveying purposes, it

may also set the standards for improvement of the management of digitisation activities.

It should be noted that the Numeric Study was a new, even ground breaking initiative. It cannot be expected to have established an EU-wide understanding of relevant definitions and surveying methodology instantly. The strengthening of awareness of a common approach and shared definitions will be needed on a permanent basis in the follow-up activities to Numeric. More specifically, the SIG considered that the training of the national coordinators that are responsible for the translations and distribution of the questionnaire, was necessary to reduce the amount of misinterpretations as encountered in the responses to Numeric and thus reach wider harmonisation.

An example to illustrate this: as identified in the Numeric Study Report, the word 'digitisation' itself proved to be problematic. Numeric used the definition from the American Institute of Museum and Library Services: "the process of converting, creating and maintaining books, art works, historical documents, photos, journals etc, in electronic representation so they can be viewed via computer and other devices." This may look like an adequate definition from an authoritative institution, but responses to the Numeric survey showed that many archives and museums, for different reasons, tend to include the cataloguing of their collections in databases as part of what they call 'digitisation'. For museums digitisation has to a large degree been part of collection management. Archives create elaborate records with information on structures and relationships between collections and objects. For them, an EAD-record can be considered as a digital object that results from digitisation. The same has been observed for monuments: do we consider a digital record of a monument as digitisation, or do we only count digital reconstructions as such? In order to obtain valid statistical data about digital heritage, these kinds of definition problems need to be solved first.

4. THE DIGITAL FACTS (NETHERLANDS)

It was precisely this need for more research and tools to support the monitoring of digitisation, that made the Dutch Ministry of Education, Culture and Science (OCW) invest in a national project on digitisation intelligence, alongside Numeric. This project was called The Digital Facts (De Digitale Feiten), and was coordinated by the DEN foundation (Digitaal Erfgoed Nederland). In the first year of the project, 2008, the focus was on getting as much as possible valid data from heritage institutions to be submitted to Numeric. A project officer worked closely with an external company to create the Dutch equivalent to the Numeric survey. A lot of effort was put into assisting the institutions to compile their responses to the survey. This was quite time consuming, but it paid off. The Numeric survey addressed many topics and the questionnaire was quite extensive. As a result, several staff members from a single institution had to get involved and proper support was needed to persuade the institutions to complete the survey. In the end, over 130 institutions agreed to participate and an authoritative publication on the progress of digitisation of cultural heritage in the Netherlands could be created. Thanks to these good results and because of their commitment to this area of research, the Netherlands were asked by the European Commission to chair the Special Interest Group, SIG-STATS.

However, as in other European countries, it was felt that not all sections of the Numeric questionnaire were based on a solid methodology. The DEN Foundation identified three main areas for further research, and the Ministry of Culture agreed to invest in a continuation of the Digital Facts project in 2009. The three main areas were methods to measure 1) usage of digital heritage collections, 2) methods to calculate costs of digitisation projects and 3) methods to measure born-digital heritage collections. In 2009 three specialised project officers were responsible for setting up recommendations on these three area's for improvement of the surveying methodology.

4.1 Web statistics

The research on the usage of digital heritage collections focused on web statistics by cultural heritage institutions. Increasingly, cultural heritage institutions provide access to digitised resources on their websites and many of them present web statistics in their annual reports. Amongst other things, the statistics can show how often a website is visited, which pages are the most popular, via which pages people enter the website etc. But what methodology and tools are used to compile these statistics? Is the use of the digital collections expressed in these statistics? How reliable are the data, and is it possible to compare the statistics across institutions and over time?

The research resulted in two reports: firstly a literature survey was carried out to obtain insight in the backgrounds, feasibilities and limitations of web statistics. This led to a practical manual for the use of web statistics. One of the recommendations is to use 'visits' as the key concept in managing web statistics, not (unique) visitors or hits, as is frequently done. Secondly a report was written on the current use of web statistics by cultural heritage institutions in the Netherlands. As was expected, only a few institutions were really aware of the many pitfalls that come with web statistics and presented their data with care. To name such a pitfall: improvement of the navigation or usability of a website may result in lower numbers of hits in the statistics, but this does not mean less use of the website. It has become easier ('less clicks') for a user to find the information and this is without a doubt a qualitative improvement. By just presenting annual web statistics in a sequence, without any explanation, wrong conclusions might be drawn. As web statistics are becoming more and more accepted as an instrument for accountability towards funds or governments, it is imperative that we make better use of them.

The reports are published in Dutch, but an English summary was created by Europeana (see References).

4.2 Costs of digitisation

The research on better ways to calculate and express costs for digitisation projects led to the creation of an elaborate cost model that can be used for project budgeting. The core of the cost model was created by the Archives of the Province of Gelderland, for their own purposes. With the support of the DEN foundation, one of their staff members investigated whether the cost model could be used by other archives and, indeed, museums and other heritage institutions. The outcome was positive, and in April 2010, the Gelders Archief and DEN were able to present the fully developed cost model, accompanied by an extensive manual.

The cost model is set up as a spread sheet, in order to give the heritage institutions full flexibility to adjust the model to their own needs. The cost model is quite extensive, allowing institutions to understand better the costs of all activities that are needed to digitise a cultural heritage collection: physical analysis, transport, adding metadata, the actual digital

reproduction through scanning or photography, quality control, storage, promotion et cetera. At the moment, the model is still being tested. If the model will be accepted widely, it will not only support harmonisation of terminology on costs, it may also support the automatic exchange of benchmarking data, if institutions are willing to share their own cost calculations.

The DEN foundation hopes to present an English version of the model in 2011.

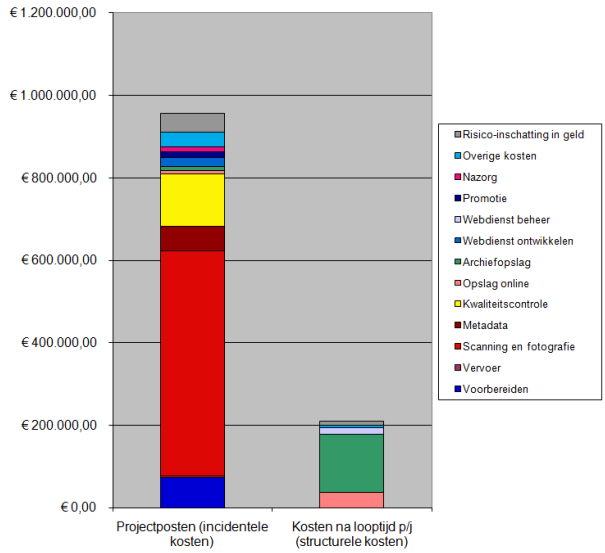

Figure 3: An example from the DEN cost model, based on a large scale news paper digitisation project. The graph projects the overall costs during and after the project lifetime, distributed across various cost categories (e.g. metadata, storage, promotion and transport).

4.3. Born digital heritage.

The third main area of research in the Digital Facts project was born digital heritage collections. This topic was not covered by Numeric, as Numeric focused on the conversion of analogue collections to digital objects. This exploratory study as part of the Digital Facts project was designed to map out specific problems of managing and measuring born-digital heritage at selected Dutch heritage institutions.

As this was really new ground, it was decided not to do a wide survey, but to focus on the heritage institutions that were considered to be pioneers with born-digital heritage materials. How do they manage and measure their collections? What problems do they encounter? In total 29 institutions participated actively. The study showed that most of their collections contain both digitised and born-digital material, that both are managed in the same system and even that it is not common to make a distinction between born-digital material and digitised material.

However, it is recognized that there are differences in acquisition, metadata and digital preservation. This is where an underlying problem surfaces. Most of the organisations only add large quantities of born-digital object types with a traditional and/or digitised counterpart to their heritage collections, such as photos, videos, audio files, e-books and e-articles. New forms of born-digital heritage, meaning objects without a traditional or digitised counterpart, are not collected or are only collected in dribs and drabs. Examples are websites, games, 3D designs or digital reconstructions.

As a result, the majority of the institutions states that interesting Dutch born-digital heritage material is being lost because it is not or not sufficiently collected, due to a lack of priority, funds, knowledge or technical facilities. There is a great need for best practices and a clear allocation of tasks among various institutions.

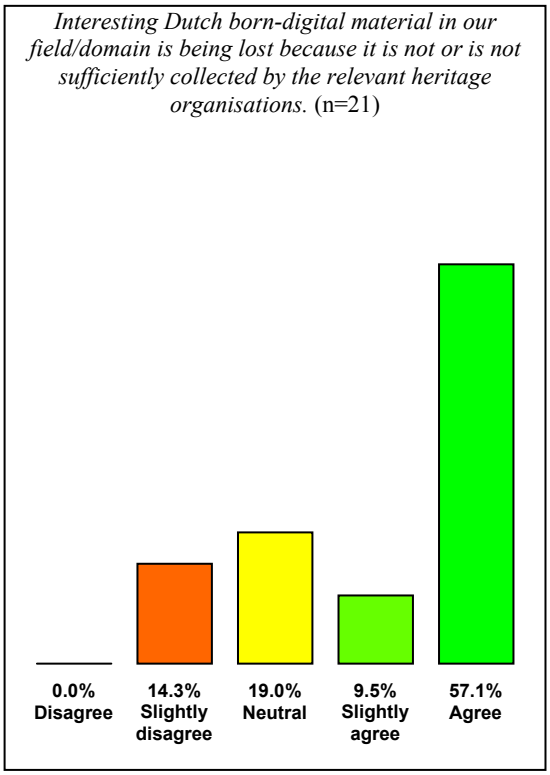

Figure 4: Graph from the report on Dutch born-digital heritage, showing that the majority of institutions agree on the fact that born digital heritage is lost due to an unclear allocation of tasks among cultural heritage institutions.

The results of these three Dutch research projects will most likely feed into a new European project that is currently in the making: ENUMERATE.

5. ENUMERATE

The Numeric project ended with the publication of the Study Report as a PDF-document. There is a wealth of information stored in the document, but the data and the graphs cannot easily be re-used or updated. SIG-STATS addressed this issue, and would very much like to see the emergence of a data repository, where statistics and other data on the digitisation of cultural heritage are not only available in static documents, but also as dynamic data that can be submitted, retrieved and visualised. Such a repository could be a valuable platform to promote networking, collaboration and knowledge-sharing about the statistical monitoring of digitisation of cultural heritage. As such, it could be an important tool to support the growth of Europeana, by providing up to date intelligence about all the content that could feed into Europeana.

In the spring of 2010, SIG-STATS decided to set the first steps towards such an data repository, by drafting a project proposal for a follow up project to Numeric. While preparing a project

proposal for the Thematic Network under the Digital Libraries theme of the EU ICT Policy Support Programme, other parties notified the SIG that they were interested in participating. These included governmental organisations or companies from Hungary, Slovenia and Spain. CollectionsTrust from the UK was invited to become the coordinator of the proposed project, to be called ENUMERATE. Together with the six original participants in the SIG-STATS, this consortium aims at a lasting transformation in the availability, quality, accuracy and relevance of statistical data about digitisation, digital preservation and online access to cultural heritage. The main objectives of ENUMERATE are:

- The development of a vibrant and sustainable European **community of practice**, connecting practitioners in statistical analysis and digital content creation and preservation and supporting the sharing of knowledge and best practices.

- The creation, promotion and development of a statistically valid **open methodology** for surveying the digitisation, use and preservation of cultural heritage materials in Member States.

- The implementation of a multi-annual programme of coordinated **surveys** based on this methodology, including wide-scale harmonized statistical data-gathering and more in-depth and analytical surveying of digitisation activities by European cultural heritage institutions.

- The creation and maintenance of an open, sustainable **data platform** to collate, analyse and promote the use of the normalized data and intelligence arising from these surveys.

At the time of writing, the ENUMERATE proposal was reviewed positively by the evaluation committee of the ICT-PSP Digital Libraries theme, but negotiations with the European Commission are still to take place. It is hoped that ENUMERATE will start in January 2011.

6. CONCLUSIONS

Both the heritage institutions and governments at various levels have a growing need for more accurate and up-to-date intelligence on the digitisation of cultural heritage. Many parties consider the transition from analogue to digital culture a landmark activity of our time, but speeding up this process requires large-scale, coordinated efforts across Europe, within Member States and between individual institutions and networks. Better data on the size, costs and use of digital heritage is needed to track impact, to identify and celebrate success, and to define policies and funding instruments to target specific issues or opportunities.

The NUMERIC project estimated that the annual value of dedicated digitisation budgets of cultural institutions in Europe added up to a total of 261 million euro (Numeric Study Report p. 69). The majority of the costs of digitisation and digital preservation are funded through public subsidy. This represents a real-terms investment on behalf of European citizens of many millions of euro every year. More quality data on the output of these digitisation efforts contribute to a better accountability to society at large.

However, there is not yet a strong tradition in gathering statistical data on digital heritage. There is no clear cut methodology to do so on a regular basis. The NUMERIC project was a ground breaking effort to set a new standard for this type of intelligence. Some satellite activities, such as SIG-STATS and the Digital Facts project in the Netherlands, contribute to the evaluation and further development of the outcomes of the NUMERIC project and to the creation of a base that is useful for future benchmarking.

The projects described in this short paper are, together with other related projects and activities, proof that there is a growing commitment to the development of methods and tools to improve our knowledge about digitisation activities and their output. If we are to create sustainable models for digital heritage services, such intelligence will prove to be crucial in the strategic decision-making by any party involved at European, national or institutional level.

7. REFERENCES

DEN, 2008. Digitaal Erfgoed Nederland, De Digitale Feiten. http://www.den.nl/ictmonitor/onderzoek/digitalefeiten. (accessed 20 August 2010)

DEN / EUROPEANA, 2009. Digitaal Erfgoed Nederland / Europeana, Web statistics of heritage institutions, Summary of a survey. http://www.den.nl/english (accessed 20 August 2010)

DEN, 2010a. Digitaal Erfgoed Nederland, Born-digital heritage materials at selected Dutch heritage organisations, an exploratory study. http://www.den.nl/english (accessed 20 August 2010)

DEN, 2010b. Digitaal Erfgoed Nederland, Rekenmodel Digitaliseringskosten http://www.den.nl/docs/20100408024532 (accessed 20 August 2010)

EGMUS, 2010. European Group on Museum Statistics. http://www.egmus.eu/ (accessed 20 August 2010)

EUROPEAN COMMISSION 2010. Competitiveness and Innovation Framework Programme (CIP), ICT Policy Support Programme, Work Programme 2010. (esp. Theme 2 Digital Libraries, Objective 2.6) http://ec.europa.eu/information_society/activities/ict_psp/documents/ict_psp_wp2010_final.pdf (accessed 20 August 2010)

NUMERIC, 2010. NUMERIC. http://www.numeric.ws/ (accessed 20 August 2010)

SIG-STATS, 2010. Follow-up to the Numeric survey on cultural heritage digitisation statistics, Recommendations from the Special Interest Group on Cultural Heritage Digitisation Statistics. http://cordis.europa.eu/fp7/ict/telearn-digicult/publications_en.html (accessed 20 August 2010)