Follow-up to the Numeric survey on cultural heritage digitisation statistics
Recommendations from the Special Interest Group on Cultural Heritage Digitisation Statistics

This document has been made publicly available to encourage the creation of project proposals for a thematic network under the CIP ICT-PSP Call for proposals 4, Theme 2 (Digital Libraries), Objective 2.6 (Statistics on cultural heritage digitisation activities) (deadline 1 June 2010).

Introduction
This document has been created by SIG-STATS, a Special Interest Group, supported by the European Commission, for the collection of statistics on the digitisation of cultural materials in Europe. The main tasks of SIG-STATS have been to review the methodology used in the Numeric study (2007-2009), and to suggest improvements for the organisation and methodology of future surveys.

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.

The Numeric report was published as a draft in May 2009, the final version was published in February 2010.

A group of six national coordinators from the Numeric Framework (Austria, Belgium, France, Germany, Hungary and the Netherlands) agreed to participate in a Special Interest Group, installed by the EU Member States Expert Group on Digitisation and Digital Preservation. In February 2010 the SIG met in Luxembourg to discuss the recommendations for a follow up to the Numeric study. This document presents the outcomes of this meeting.

Based on the recommendations in the Numeric Study Report and the individual reviews of the report by the SIG members, possible scenarios for seven specific topics were discussed at the meeting:

1. Survey design: principles to structure (or breakdown) 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 analog heritage to the digital representations
5. Calculation of costs
6. Measuring use and access: valid and practical means of measuring access to and use of digitised heritage
7. The framework: organization and implementation of the survey across EU27 countries

The next paragraphs will present the recommendations of the SIG on all these issues.

1. Survey design: principles to structure (or breakdown) the questionnaire

Review
Although close to 800 institutions responded to the Numeric survey in total, in many countries the amount of respondents was considered rather low, which jeopardises the quality of the data. The SIG thinks it is crucial to reconsider the way the questionnaire was structured. The form of the original Numeric questionnaire needs improvement, to get more and better responses. The main issues, that have also been recognised in the Study Report itself, were:

a. The Numeric questionnaire was rather extensive, which resulted in quite a heavy burden on the institutions to fill it in properly. Cultural heritage institutions are frequently asked to contribute to regional, national or international surveys by research projects or policy makers. An extensive questionnaire like this one may just be a bit too much for many institutions.

b. The extensiveness of the survey resulted from the many topics that were addressed in the questionnaire. As a result, several staff members from a single institution were needed to contribute to the survey. This may not only have affected the overall quality of the contribution, it also made the communication between the study team and the respondents more complex.

c. Some questions, esp. the ones on a detailed level, had quite a low response rate. There are several explanations for this:

- the answer to the question is not at hand, and additional research was needed to provide the answer. It is believed that this is the reason why many institutions were not able to provide financial data, as these institutions did not have a specified budget for digitisation. In fact, some institutions had observed how difficult it was to distinguish between (general) IT resources and the proportion of the financial allocation devoted to digitisation;
- the question could be too technical for the staff member responsible for the questionnaire (e.g. the values for scanning resolution);
- the institution may not be willing to share confidential information with the survey team (e.g. information about contractors or service providers);
- a low response to a particular question may reflect the fact that the question is not relevant to all types of institution.

Recommendations
The SIG firmly believes that the Numeric Study should not only be about compiling facts 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 more precise information about the size, costs and use of their digital collections. Numeric should not be only about short term statistics, but also about long term accountability and performance indicators. It will take a few years of research in order to acquire useful benchmarking data.

The SIG recommends a hybrid approach, which on the one hand will not compromise the original goals of the 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 SIG recommends to split up the questionnaire into a short version, which can be sent to all cultural heritage institutions in Europe and a more elaborate version, to obtain more detailed information from a selected sample.

1. The short questionnaire (or 'core questionnaire'): the three high level questions of the short questionnaire refer to costs (annual investment per institution in digitisation), size of the digital collection and use of the digital collections and services of the institutions. These data can be used by the institutions in their annual reports. Therefore, it should be considered to

have an annual update of this short questionnaire (as is currently put foward by the Conference of European National Librarians). The short questionnaire should also address the typology of the institution (e.g. to what subdomain does it belong and what are the total size and budget).

2. The full questionnaire: This questionnaire should handle the three main topics (size, costs and use) in more detail, as was done in the original Numeric questionnaire. The SIG sees two options to go forward and lessen the burden for the institutions to participate:

    a. an incremental approach: start with a medium sized singular questionnaire, provide proper training and support and extend the survey over the years

    b. a modular approach: breakdown the questionnaire into three thematic questionnaires, that can be sent to a more targeted sample of institutions (e.g. based on the outcomes of the short questionnaire; if an institution is not able to provide information about costs of digitisation in the short questionnaire, there is no point in sending it the full questionnaire).

With the short questionnaire, a representative overview can be given of digitisation of cultural heritage in Europe; with the full questionnaire more qualitative information can be acquired to support benchmarking, quality assurance etc.

2. Defining the survey sample

Review
The Numeric Study Report presents the data from "a representative mix (at European level) of archives, museums, libraries, and institutions specialising in audio-visual and other heritage collections." This representative mix was achieved by applying the criterion of 'Relevancy'. Relevancy of institutions is used in the Numeric Study "in the sense that digitisation of their collections would significantly enhance access to the country's cultural heritage." Relevancy was introduced to make international comparisons possible, no matter how many (or few) heritage institutions there are in a particular country.

The SIG confirms that, as is obvious from the Numeric Study report itself, the composition of the survey sample based on 'relevancy' has been one of the most problematic issues of the survey. The report dedicates several pages to the different ways the national coordinators identified 'relevant institutions' in their contributions.

There are several reasons why this point of departure for choosing the sample is problematic. As the Study Report itself proves, the definition of 'relevancy' is open to various (mis)interpretations. Some put the focus on a 'country's cultural heritage' (implying a focus on national institutions). Others focused on 'enhance access' (implying a focus on institutions that are digitally advanced and connected to the web). More problematic in general, is the fact that relevancy introduced the subjective concept of 'cultural value' to the survey methodology. In one country a museum of thimbles may not be taken seriously, in another this could very well be an important heritage collection. It is not feasible for national coordinators to make these kinds of value assessments while selecting a sample. There has obviously been no agreement or clarity about the sample method to be implemented jointly.

However, the principle of guaranteeing the quality of international comparisons of data is, of course, a valid one. The SIG would like to emphasize that the key concept for the study should be 'representativeness', not 'relevancy'. In the Numeric Study, representativeness is coined as a term as well, but only to designate the sample within the set of relevant institutions. Representativeness is needed in order to live up to one of the original goals of the survey: measuring the progress in digitisation of cultural heritage in Europe.

Recommendations
The SIG would like to propose the following approach, where the qualitative assessment of cultural value is abandoned:

a. the top level is what is called the "cultural heritage domain". Every institution that belongs to this domain, whether publicly or privately funded, whether actively involved with digitisation or not, may contribute to the Study;

b. the cultural heritage domain can be broken down into subdomains, or types of institutions. The list with types of institutions as used in the Numeric Study Report can serve as a good starting point.

Archive/records office
Audio-visual or film institute
Broadcasting institute
Museum of art, archæo, hist
Museum of science, tech, ethn
Other type of museum
National library
Higher education library
Public library
Special or other type of library
Other type of organisation

Some alterations may be needed for a new survey, but this needs further discussion. For instance, more hybrid institutions are emerging (e.g. 'historical centres', a combination of a museum and an archive); not all public libraries are considered to be 'memory institutions', as they do not intend to keep their collections intact over longer periods of time; and institutions dealing with non-tangible heritage, monuments or landscape preservation may need their own designation instead of being grouped together as 'Other type'.

c. Once the list of subdomains has been finalised at the central level, it is up to the national coordinator to decide on the size and composition of each subdomain. There is no EU-wide consensus on what each subdomain exactly looks like, and the SIG believes that this cannot be achieved overnight, as the national subdomains may have been created and administered over decades or even centuries. Also, in one country, there may be a good administration of privately funded heritage institutions, in others, the national coordinator may not have this kind of information at hand. Overall, the SIG recommends to put the priority with the (fully or partially) publicly funded institutions.

d. As discussed above, the SIG recommends to have two types of questionnaires, a very short one, that will focus on the basic questions regarding size, costs and use of digital heritage collections and a more extensive questionnaire to get more in-depth information. The SIG recommends:

1. that the short questionnaire will be sent to <u>all</u> institutions that belong to the cultural heritage domain. This is the best option to get a true representative view (within statistical boundaries) of the current status of digitisation of cultural heritage in Europe. By including a question to which subdomain the institution belongs according to its own criteria, the national coordinator can refine the composition of subdomains in the longer run.

2. that the full questionnaire will be sent to a selection of institutions within each subdomain. A precise sample method needs to be determined once the exact form of the full questionnaire (incremental or modular) has been chosen. In general, the SIG thinks that a modified version of the Numeric instrument to calculate the representative sample is helpful in this respect.

3. Definitions

Review
The Numeric project did some excellent work on definitions on digitisation of cultural heritage. It is crucial that all parties involved adopt common definitions in order to get good quality data from each participant. Careful attention has been paid by the Numeric team to existing definitions from various sources to support the cultural institutions that contributed to the survey.

Nevertheless, there is some room for improvement. The SIG distinguishes two sides to this topic: the use of terminology in the questionnaire that can be understood by all cultural institutions ('jargon') and the multilingual aspects of the terminology, resulting from the translation of the questionnaire into the EU27 languagues.

The SIG would like to put forward some generic and some specific comments on the terminology used in the Numeric survey.

Generic recommendations
a. Since the Numeric Study was a new, even ground breaking initiative, it cannot be expected to have established an EU-wide understanding of relevant definitions instantly. The strengthening of awareness of the importance of common definitions will be needed on a permanent basis in the follow up activities to Numeric. More specifically, a training of the national coordinators that are responsible for the translations of the questionnaire, is considered necessary to reduce the amount of misinterpretations and reach wider harmonisation.

b. The three main topics of the survey (size, costs and usage of digital heritage) are the results of complex sets of activities and procedures. The SIG recommends, in addition to the study of existing definitions, to study documents and tools that reflect the overall 'digital workflows' in the cultural institutions. By understanding the workflows, more precise definitions can be provided, for instance on the creation of digital collections, the budgets for staff involved in digitisation projects and the types of access to digital heritage services. It is obvious that further development of the definitions needs to take place in direct dialogue with the institutions themselves.

Specific recommendations
a. As was already identified in the Numeric Study Report, the word 'digitisation' is itself problematic. Numeric used the definition from the IMLS: "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." As such, this is an adequate definition from an authoritative institution, albeit perhaps too much from the point of view of libraries. 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 eloborate 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?

As digitisation is not a phenomenon that is restricted to cultural heritage, the SIG recommends to use the more generic Wordnet definition for 'digitisation' as a general starting point: "conversion of analog information into digital information", and from there make explicit distinctions between 'digital descriptions' (or 'metadata') and 'digital reproductions' (or: 'representations') in future surveys.

b. Another definition issue relates to digitisation costs. The SIG observed that the Numeric Study Report shows big differences in costs for digitisation of specific collections, e.g. audiovisual collections. It is possible that these differences result from different quality criteria, but it may also be that some institutions included specific activities in the costs, and others did not. This is not clear, and the SIG considers this a matter of proper definitions. The SIG recommends to identify and analyse existing cost models for digitisation, and use the categories in these models for more precise

guidelines for supplying information about digitisation costs. The Numeric Study Report also suggested the use of a checklists, e.g. of common digitisation processes and activities (including web development) and of types of staff members involved in digitisation. The SIG considers this a good starting point for further development of definitions on digitisation costs. (more information on costs in par. 7).

d. Future surveys should try to use a single reference period as much as possible. In the Numeric questionnaire, various reference periods were applied (2007, 2007/2008, 2008, 2008/2009).

e. In the Numeric Report, the number of digitisation projects are equalled to the number of digital collections that result from it. This is not a valid comparison. Not all projects are dedicated to a single collection, and not every digital collection is created in a single digitisation project.

4. Input-output measures

Review
One of the objectives of the Numeric survey was to measure progress (growth) in the digitisation of European cultural heritage collections. Uniform and explicit measures are necessary to get comparable results across different countries. The Numeric questionnaire contained a fairly extensive table of object types that could possibly be present in a collection. The table was designed to measure the size of analogue collections (input) and the size of realized and planned digitisation of these collections (output).

There were two main problems in this respect. 1. The willingness to fill out the table turned out to be rather low. The rule seems to be: the more detailed the table, the poorer the response. 2. Although most input-output measures posed no problems as such, a few categories turned out to be troublesome, in particular: archival records, newspapers, and monuments.

Another point that needs closer attention in the follow up is born-digital heritage. At the time of reflecting and composing the original Numeric questionnaire born-digital objects were still thought of as a separate category of cultural heritage objects. The survey was about digitisation in the strict sense and it was decided to leave born-digital objects out. Today, e.g. in big audio-visual institutions, the difference between these two types of objects (digitised and digital by origin) is becoming irrelevant. It needs to be questioned whether born-digital heritage can be left out again in the future.

Recommendations
Given the recommendation to split up the overall questionnaire, it needs to be decided how input/output measures are included in both the core questionnaire and the full version. The SIG recommends the following approach.

Core questionnaire: the SIG recommends a rough figure of the size and growth of collections across Europe will suffice. Therefore the percentage of the entire collection per institution that *has already been digitized* will be measured here. And to put this simple measure into perspective the core questionnaire will also canvass both the percentage of the entire collection(s) that *need not be digitized* as well as the percentage that *still needs to be digitized*. The SIG recommends against the use of the broad categories "images, text, audio, video" as an alternative to the lengthy table in the full questionnaire. There are just too many types of hybrid digital objects to make this a useful distinction.

Full questionnaire: it is necessary to ask for more detailed input-output measures using a more elaborate table. In general, the table used in Numeric, serves as a good startingpoint. It will be a task for the Thematic Network to consider solutions for the more troublesome object types. Additional research is needed to validate the objects and measurement units in the current table. The SIG thinks that a quick analysis of object types in some of the major collection management systems could provide a quick win in that respect.

In the SIG there was no full consensus about the incorporation of born-digital materials in future surveys. Preservation problems will compel the heritage sector to formulate new policies, and as such it is useful to start gathering information on this type of collection. However, as the Numeric questionnaire was already quite complex as it was, introducing born-digital heritage can only make the questionnaire more complex. All in all, the SIG recommends to use the full questionnaire for some exploratory research on born-digital issues. A decision on the extent of this research is left to the Thematic Network.

5. Calculation of costs

Review
Substantial variance was found in the Numeric survey concerning the recorded costs of digitisation. Differences even occurred across projects within the same institution. As a consequence, one cannot determine one proper price for the digitisation of a certain type of collection. As suggested above (see the paragraph on definitions), the way questions about expenditure were phrased may have resulted in too much ambiguity. Another explanation for these differences could lie in the inclusion or exclusion of 'invisible costs', such as staffing costs, costs of making policy or project plans, etc. The SIG came to the conclusion that it is not justified to make generalisations about digitisation costs solely based on the responses in the Numeric survey. Calculating costs for digitisation projects is a very complicated issue, yet for the management of institutions, policy makers and funding agencies a better insight in these costs would be profitable.

Recommendations
In line with the recommendations for surveying the growth of collections and measuring the use being made of digital cultural heritage, the SIG recommends to make a distinction between a few global specifications of costs that can be gathered with the core questionnaire, and more detailed data from the full survey, or indeed a separate survey on costs.

For the core questionnaire the SIG recommends to focus on a few major cost categories (e.g. staff, equipment) that add up to the total expenditure on digitisation. The Numeric Study Report suggests to ask for the total expenditure in two consecutive years, last year and current year. The SIG believes it will be better to exclude questions about the current year, and focus on the two previous years. The report also suggests to ask for the costs of planned digitisation projects. The SIG expects that only few institutions will be able to provide this kind of data. As suggested earlier, further research on existing cost models will be useful to determine the major cost categories.

In the full questionnaire a more thorough breakdown of costs would be desirable, but additional research is needed here. Analysis of existing cost models (e.g. JISC, Prestospace, DEN) must be taken into consideration in any follow up activity. For some time to come, it seems inevitable that a free text field must be added to the questionnaire so the institutions can explain the data about costs and budget as provided.

6. Measuring use and access

Review
A highly problematic issue identified in the Numeric Study Report are the figures collected on use of and access to digital heritage collections. Although all parties agree that measuring use and access is important, the Numeric survey - and other research as well - made clear that there is still a long way to go before trustworthy measures of use and access across institutions, let alone across countries, can be made. Detailed insights into the behaviour and preferences of visitors can in principle be distilled from the logs of online and offline collection databases, but there are many pitfalls in these measurements,

most having to do with a lack of standards. Web statistics analysis is still in its infancy, not just in cultural heritage, but in any area.

Recommendations
The SIG agreed that some measure of usage and access is necessary, both in the core questionnaire and the full version, but recommends to keep the measurements as simple as possible, and complement the questionnaire with some additional qualitative research.

The core questionnaire should yield a general indicator of what percentage of an institution's digital collection is available, both offline and online. Added to this should be a question on whether the use of the digital collection is measured, and if so, what methodology of measurement is chosen. The information gathered this way can be used to determine the target group for the full questionnaire and/or the qualitative research.

In the full questionnaire more detailed information on use statistics and access rights can be investigated, in combination with a *qualitative analysis* of survey results. The latter can be limited to a few case studies. Over time it may become possible to get more reliable quantitative results from a larger group of contributing institutions, depending on the adoption of common standards for (web) statistics.

7. The framework (the organization of future surveys)

Review
The Numeric Survey succeeded in establishing a European wide framework to get respresentatives from all EU27 countries involved. The framework consists of the following stakeholders:

- European Commission [giving support]
- national ministries [encouraging through MSEG; managing national parties]
- central (national) statistical divisions/agencies [survey setup; contextual data; data analysis]
- other institutions that are already involved in comparable surveys [alignment]
- facilitator on EU level [coordination; organizing technical support and services]
- national coordinators [national coordination; informing parties; enabling data (re-)use]
- cultural heritage institutions [providing data; benchmarking]
- external experts [consultancy]

The Numeric Study Report stressed the importance of sound procedures for the organisation and implementation of the entire survey, as so many stakeholders are involved. Ideally the number of parties involved in collecting data should be kept to a minimum, but the “middleware” cannot be left out, since the national coordinators have their indispensable networks and are necessary to overcome possible language problems.

Building on the results of the Numeric Survey, three main topics need to be considered to strenghten the framework: a. the stakeholders/actors involved, their roles and the support they need and get; b. ways to strengthen the motivation of institutions to participate; c. how to organise proper guidance during the survey.

Recommendations

a. Stakeholders
The SIG has some worries on the workload involved for some stakeholders, in particular the national coordinators. The SIG recommends three action lines:
1. Make the workflow of the Numeric survey (or its follow up) explicitly visible, showing all the roles and relationships between the stakeholders. This will make the workload clear and expectations about contributions to the framework can be realistic.

2. Create a list of the responsibilities for the national coordinators, who are key persons in the framework. A national coordinator may rely on other domain experts in his/her country for specific task, but that can only be handled per country (by the coordinator in consultation with the MSEG representative).
3. Involve as much as possible other (national) parties that run surveys on the same or related topics. This may lessen the burden on the individual institutions, as data can be acquired from previous surveys. Existing surveys can also provide means to assess the quality of newly gathered data.

b. Strengthening motivation
Low response rates, not only on the EU-level but also at the national levels, can be fatal for future surveys. The SIG recommends to set clear targets in order to decide at what level/response rate survey results can be declared valid/representative. The Numeric Study Report can serve as the minimal point of departure. The report noted "a high degree of cynicism regarding the benefits that [...] contribution of data can make to gaining resources for [...] activities." Targeted actions to improve participation are therefore necessary. The SIG thinks that providing simple tools to support the gathering of information and providing support are valid ways to increase involvement. Additional research is needed to conceive appropriate stimuli to participate.

c. Guidance
Although the SIG thinks that the Numeric approach to implementing the framework was sound and should be kept intact as much as possible, some better guidance procedures will be necessary. There are two leads here: on the one hand the national coordinators must be guided more intensely through the process of implementing the survey and chasing survey response; on the other hand the institutions should get better guidance in filling in the questionnaires. The Numeric Study Report contains a Guide to Implementing the Numeric Framework. Apart from this guidance on paper, the SIG recommends to offer live help, e.g. in the form of training facilities and opportunities to meet and discuss issues raised during work on the survey.

The SIG realises that this live support may get complicated when the original Numeric survey will be broken down into a core questionnaire and full questionnaire (which may end up in three thematic questionnaires if the modular approach will be applied). The only conclusion can be that proper guidance should be a core activity of any follow up survey.

On behalf of the SIG-STATS,

Austria: Federal Ministry for Education, the Arts and Culture , Irene Hyna
Belgium: FARO, Jeroen Walterus
France: Ministère de la Culture et de la communication, Sonia Zillhardt
Germany:Stiftung Preussischer Kulturbesitz, Institute for Museum Research / EGMUS, Monika Hagedorn-Saupe
Hungary: Oktatási és Kulturális Minisztérium, Gábor Veres
Netherlands: The DEN Foundation, Marco de Niet & Gerhard Jan Nauta

9 April 2010