



Grant Agreement 270939

ENUMERATE

Overview of Harmonisation Tools

Deliverable number	<i>D2.2</i>
Dissemination level	<i>Public</i>
Delivery date	<i>September 2011</i>
Status	<i>Final</i>
Author(s)	<i>Sjoerd Bakker, Marco de Niet, and Gerhard Jan Nauta (DEN)</i>



This project is funded under the
ICT Policy Support Programme part of the
Competitiveness and Innovation Framework Programme.

Contents

1	EXECUTIVE SUMMARY	4
2	INTRODUCTION: HARMONISATION AND VALIDATION.....	4
3	TOOLS	5
3.1	Tools	5
3.1.1	NUMERIC Project.....	5
3.1.2	NUMERIC Terminology	5
3.2	Cost models	11
3.2.1	SIG-STATS recommendations.....	11
3.3	Collection type analysis	14
3.4	Guidelines for web statistics	15
4	CONCLUSIONS	16

1 Executive Summary

This document is an overview of validation and harmonisation tools concerning terminology, digitisation costs, collection types and web statistics. New tools will be added as they become available.

2 Introduction: Harmonisation and Validation

This deliverable compiles and describes tools and other instruments that are beneficial to the entire ENUMERATE thematic network. These tools have been designed to achieve standardisation or at least harmonisation in specific areas in the cultural heritage sector. In the context of ENUMERATE, the most important tools concern:

1. **Terminology:** These tools enhance the understanding of phrases or even jargon that is used in the ENUMERATE Core Survey.
2. **Costs of digitisation:** These tools help to assess components of cost structures of digitisation activities.
3. **Collection type analysis:** These tools are created to make uniform typologies of cultural heritage collections as kept by memory institutions.
4. **Web statistics:** These tools help to use and interpret web statistics in a valid way.

This deliverable will assist the ENUMERATE consortium to ameliorate the questions in the ENUMERATE surveys, by weighing the words and ideas reflected in the survey against external sources.

Additionally, it will help train the National Coordinators, preparing the rollout of the survey in their own countries.

Finally, it will enable the institutions that contribute statistical data to the thematic network to check the validity of their data by implementing one or more of the tools described in this document.

This document will be updated regularly during the lifespan of the ENUMERATE Thematic Network project.

3 Tools

This deliverable compiles and describes tools and other instruments that are beneficial to the entire ENUMERATE thematic network. These tools have been designed to achieve standardisation or at least harmonisation in specific areas in the cultural heritage sector. In the context of ENUMERATE, the most important tools concern:

3.1 Tools

3.1.1 NUMERIC Project

In the NUMERIC project, which was carried out from 2007 until 2009, an effort was made to establish clear classifications and definitions in order to limit as much as possible the ambiguity of the survey questions. To this effect an extensive list was compiled with relevant definitions that were available to be adopted as standard definitions in surveys. These definitions were tested in a 'pathfinder' survey, which was filled out by 60 respondents from various EU and non-EU libraries, museums and archives. The comments that arose in this test survey were addressed and, if applicable, processed in the creation of the ultimate survey.

The aim of ENUMERATE is "not to reinvent the wheel," and, consequently, the NUMERIC definitions will be of use in constructing the definitions for the present survey. Criticism that arose after the completion of NUMERIC and the recommendations of SIG-STATS will be incorporated in the ENUMERATE survey.

The recommendations made by the SIG-STATS workgroup are summarised in the next section.

3.1.2 NUMERIC Terminology

Overall, the SIG-STATS workgroup deemed the work that was done in determining the NUMERIC definitions to be excellent. However, there is room for improvements and the SIG-STATS workgroup has split these up into two sides, namely: the use of jargon, which can be understood regardless of geographical location, and "the multilingual aspects of the terminology." For the resolutions of the 'flaws' in the NUMERIC survey SIG-STATS proposed some generic and some specific recommendations.

Generic recommendations

- The awareness of the importance of common definitions needs to be strengthened. This can be achieved by training the national coordinators who are responsible for the translation of the survey.
- In NUMERIC, existing definitions have been studied. For more accurate definitions SIG-STATS proposes the studying of the 'digital workflows' of cultural institutions. The development of the definitions needs to be done in collaboration with the cultural institutions. Digital workflows will be analysed as a part of the Thematic Survey in the second year of the ENUMERATE project.

Specific recommendations

- The 'problematic' definition of 'digitisation' that was used in NUMERIC was deemed to be 'adequate' but "*too much from the point of view of libraries.*" SIG-STATS recommended using the Wordnet definition of digitisation ("*conversion of analogue information into digital information*") as a starting point and making use of clear distinctions between 'digital descriptions' (metadata) and 'digital reproductions' (representations).
- The costs of digitisation projects varied widely in NUMERIC. SIG-STATS subscribed this to a lack of proper definitions for the different categories of costs. It is recommended that existing cost models are studied and their categories used for the creation of proper definitions of digitisations costs. The use of checklists, as happened in NUMERIC, is considered to be a good starting point.
- "*Future surveys should try to use a single reference period as much as possible.*"
- "*In the NUMERIC Report, the number of digitisation projects are equalled to the number of digital collections that result from is. 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.*"

Below are some existing terminology lists that can be employed by respondents when filling out the ENUMERATE surveys. The selection is based on the presence of a substantial number of terms dealing with digitisation accompanied by a (more or less) extensive description or explanation.

Title	<i>Vocabulary of Semantic Categories</i>
Developing or managing organisation	Zinaida Manzuch for the NUMERIC project
Available since	2007
Type of institution	All
Description	<p>Annex 3 of the NUMERIC desk research finding report contains a vocabulary that is composed in the following manner:</p> <p><i>“Dictionary is composed of terms that were created by two different techniques:</i></p> <p><i>1) Part of categories are constructed – these categories are supplied with definitions;</i></p> <p><i>2) Other part of categories are recorded repetitive names of different phenomena provided in surveys, they are not followed by the definition.”</i></p> <p>The annex can be found on pages 65-69.</p>
Link	http://www.numeric.ws/uploaded_files/NUMERIC_Desktop_Research_on_Digitisation_Studies26102007114936.pdf

Title	<i>A Glossary of Archival and Records Terminology</i>
Developing or managing organisation	The Society of American Archivists (SAA)
Available since	2005
Type of institution	Archive
Description	<p><i>“This glossary is based primarily on archival literature in the United States and Canada, in that order. In a few instances, terms, definitions, and citations from other English-speaking communities are included when relevant. This glossary includes terms that relate to the types of records that someone is likely to encounter when reading archival literature or when working with a fairly typical collection of records and it emphasizes terms relating to electronic records. It also incorporates terms from the literature of preservation, law, and micrographics, as well as common form and genre terms from architectural and technical drawings, motion picture and video, photography, and sound recording. It includes some words that are no longer in common use, but which are useful when reading older literature; for example, Spindex. The glossary does not include many words specific to affiliated professions, such as rare books or printing.”</i></p> <p>(Source: http://www.archivists.org/glossary/Introduction.asp)</p>
Link	http://www.archivists.org/glossary/index.asp

Title	<i>Online Dictionary for Library and Information Science</i>
Developing or managing organisation	Reitz, Joan M.
Available since	2004
Type of institution	Library
Description	A dictionary with 4000 entries from the fields of publishing, printing, literature and computer science that the author determined to be of use to library professionals.
Link	http://www.abc-clio.com/ODLIS/odlis_A.aspx

Title	<i>International Guidelines for Museum Object Information: The CIDOC Information Categories</i>
Developing or managing organisation	The International Committee for Documentation of the International Council of Museums (ICOM)
Available since	1995
Type of institution	Museums
Description	<p><i>“The International Guidelines for Museum Object Information: The CIDOC Information Categories is a description of the Information Categories that can be used when developing records about the objects in museum collections. The Guidelines can be adopted by an individual museum, national documentation organization, or system developer, as the basis for a working museum documentation system.</i></p> <p><i>The Guidelines incorporate the following elements:</i></p> <ul style="list-style-type: none"> • <i>a definition of the Information Categories that should be used when recording details about objects;</i> • <i>an outline of the format rules and conventions governing how information is entered in these categories;</i> • <i>comments on the terminology that can be used in these categories.”</i> <p>(source: http://cidoc.mediahost.org/guidelines1995.pdf)</p>
Link	http://cidoc.mediahost.org/guidelines1995.pdf

Title	<i>ISO 2789 Information and documentation -- International library statistics</i>
Developing or managing organisation	International Organization for Standardization
Available since	2006
Type of institution	[no information]
Description	<p>Abstract:</p> <p>ISO 2789:2006 specifies rules for the library and information services community on the collection and reporting of statistics</p> <ul style="list-style-type: none"> • for the purposes of international reporting, • to ensure conformity between countries for those statistical measures that are frequently used by library managers but do not qualify for international reporting, • to encourage good practice in the use of statistics for the management of library and information services, and • to specify data provision required by ISO 11620.
Link	http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=39181

Title	<i>ISO 5127 Information and documentation -- Vocabulary</i>
Developing or managing organisation	International Organization for Standardization
Available since	2001
Type of institution	All
Description	[no information]
Link	http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=33636

Title	Wordnet
Developing or managing organisation	Princeton University
Available since	2010 (?)
Type of institution	All
Description	<p>The SIG-Stats workgroup has recommended that follow-up project to NUMERIC, like ENUMERATE, use the Wordnet definition of digitisation. The contents of Wordnet itself, however, are not limited to definitions that are related to relevant subjects.</p> <p><i>“WordNet superficially resembles a thesaurus, in that it groups words together based on their meanings. However, there are some important distinctions. First, WordNet interlinks not just word forms – strings of letters – but specific senses of words. As a result, words that are found in close proximity to one another in the network are semantically disambiguated. Second, WordNet labels the semantic relations among words, whereas the grouping of words in a thesaurus does not follow any explicit pattern other than meaning similarity.”</i></p> <p>(Source: http://wordnet.princeton.edu/wordnet/)</p>
Link	http://wordnet.princeton.edu/wordnet/

Title	Guidelines for the Preservation of Digital Heritage
Developing or managing organisation	UNESCO
Available since	2003
Type of institution	All
Description	Of particular interest to the ENUMERATE project are page 20, which contains definitions of the terms digital preservation, digital materials, preservation programme and presentation/re-presentation; as well as page 157 to 159 which contains a glossary explaining the term as used in the guidelines.
Link	http://unesdoc.unesco.org/images/0013/001300/130071e.pdf

Title	<i>Digital Heritage Collections – Glossary of Terms</i>
Developing or managing organisation	Collections Council of Australia
Available since	2008
Type of institution	All
Description	A glossary of term aimed solely at digital collections.
Link	http://www.collectionscouncil.com.au/Portals/0/Digital%20Heritage%20Collections%20-%20Glossary%20of%20terms.pdf

3.2 Cost models

It is in the ENUMERATE project's benefit if respondents to the questionnaires supply uniform data. One area where it is likely that the individual interpretation of an institution will result in widely differing data is the calculation of digitization costs. The SIG-STATS workgroups has made some recommendations directed at resolving some of the issues.

3.2.1 SIG-STATS recommendations

The SIG-STATS workgroup has made recommendations for both the core questionnaire and the full questionnaire. For both questionnaires, however, it is suggested that existing cost models (JISC, Prestospace, DEN etc.) are studied.

The questions related to digitization costs in the core questionnaire are to focus on some major cost categories. Instead of asking for the expenditure in the last and current year, as was done in NUMERIC, the SIG-STATS workgroup recommends focusing on the two previous years, as the data for those years is presumably no longer subject to change.

The questions in the full questionnaire will benefit from a more thorough breakdown of costs. To achieve this research into existing cost models is needed.

Below are some existing tools that can be employed by institutions to calculate their digitisation costs.

Title	Rekenmodel digitaliseringskosten [Calculation model digitalisation costs]
Developing or managing organisation	Stichting Digitaal Erfgoed Nederland (DEN) in cooperation with Erfgoed Nederland.
Available since	2009
Background	The development of this calculation model was based on talks with, and input from, the heritage field. It was inspired by the 2008 calculation model of the archives of the province of Gelderland (Gelders Archief).
Description	This calculation model is created as an Excel file and it features a comprehensive set of cost categories covering a great variety of costs, from the inception to the completion of a digitisation project. Cost categories include: preparation, transport, scanning and photography, metadata, quality control, online storage, promotion and after care. Per category, a couple of options are given to calculate the cost of the respective cost category.
Link	http://www.den.nl/standaard/202/

Title	<i>Preservation Project Cost Calculator</i>
Developing or managing organisation	PrestoSpace project
Available since	2007
Background	One of a set of calculators developed as part of the PrestoSpace project.
Description	<p><i>“The SAM analysis tools allow an archive to estimate the costs involved in a digital preservation project. They address different approaches to storage management and the costs involved in the digital preservation exercise itself. A further tool estimates the potential return on investment for various business models. Taken together they provide the basis for costing and putting forward a funding case for a preservation project.</i></p> <p><i>The tools require you to enter data relevant to your archive. In the early stages this may be difficult to do so they can also provide typical data as the basis of an initial calculation. You can refine the data at a later stage. The tools are backed by tutorials which provide the background information you need to provide the information for the analysis.”</i></p> <p>(source: http://digitalpreservation.ssl.co.uk/about/56/59.html)</p>
Link	http://digitalpreservation.ssl.co.uk/hosted/d13.2/newcalc.php

Title	<i>Digitisation Costs Calculator</i>
Developing or managing organisation	Collections Trust (author Nick Poole)
Available since	2011
Background	Produced as part of the EU-funded ' <i>Study to Assess the Cost of Digitising Europe's Cultural Heritage</i> '
Description	This simple calculator, presented in an Excel file, features separate fields for museum, libraries, archives and audio-visual institutions. When the number of materials to be digitised is entered an estimated price range is given for the project.
Link	http://www.collectionslink.org.uk/discover/sustaining-digital/723-digitisation-costs-calculator

Title	<i>IMPACT Digitisation Cost Estimator</i>
Developing or managing organisation	Gottingen State and University Library for The IMPACT Project
Available since	2011 (pilot)
Background	Currently a pilot tool that was developed as part of the IMPACT project that is aimed at improving access to historical texts and taking away barriers that stand in the way of mass digitisation of European cultural heritage.
Description	A single sheet Excel file that can be filled out to find out what the estimated cost for an institution's digitisation project will be. All costs can be entered manually and the calculator will determine the total price and the price per page. The calculator is intended for institutions wishing to digitise books/texts.
Link	http://www.impact-project.eu/taa/strat/pilot-tools/

Title	<i>DiCoMo: An Algorithm Based Method to Estimate Digitization Costs in Digital Libraries</i>
Developing or managing organisation	A. Bia (Miguel Hernández University, Spain) and J. Gómez (University of Alicante, Spain)
Available since	2005
Background	
Description	<p>Abstract:</p> <p>The estimate of digitisation costs is a very difficult task. It is difficult to make exact predictions due to the great quantity of unknown factors. However, digitization projects need to have a precise idea of the economic costs and the times involved in the development of their contents. The common practice when we start digitizing a new collection is to set a schedule, and a firm commitment to fulfil it (both in terms of cost and deadlines), even before the actual digitization work starts. As it happens with software development projects, incorrect estimates produce delays and cause costs overdrafts.</p> <p>Based on methods used in Software Engineering for software development cost prediction like COCOMO and Function Points, and using historical data gathered during five years at the Miguel de Cervantes Digital Library, during the digitization of more than 12.000 books, we have developed a method for time and cost estimates named DiCoMo (Digitization Costs Model) for digital content production in general. This method can be adapted to different production processes, like the production of digital XML or HTML texts using scanning and OCR, and undergoing human proofreading and error correction, or for the production of digital facsimiles (scanning without OCR). The accuracy of the estimates improve with time, since the algorithms can be optimized by making adjustments based on historical data gathered from previous tasks.</p>
Link	http://www.springerlink.com/content/f3l24618np58427n/

3.3 Collection type analysis

In the NUMERIC project, materials which are held in the collections of heritage institutions were classified on the basis of a study of 32 digitisation reports.¹ The SIG-STATS workgroup deemed the ultimate list of classifications to be a good starting point for subsequent surveys, but did offer some recommendations concerning specific classifications, namely: archival records, newspapers and monuments; which need to be classified more clearly, and born-digital objects; which were not included in the NUMERIC survey but will be in ENUMERATE.

Title	<i>Classifications of materials held in collections</i>
Developing or managing organisation	NUMERIC project
Available since	2009
Background	All
Description	The list of classifications as developed and employed by the NUMERIC project. These classifications were drawn up after the examination of 32 digitisation reports. The table in the final report that contains the classifications also contains the units in which they are to be measured. The table can be found on page 17 of the final Numeric study report.
Link	http://cordis.europa.eu/fp7/ict/telearn-digicult/numeric-study_en.pdf

Title	<i>Defining “Born Digital”</i>
Developing or managing organisation	Online Computer Library Center (OCLC)
Available since	2010
Background	All
Description	“Born-digital resources are items created and managed in digital form” is the definition that has been used in compiling this list of types of born-digital materials. For each type a definition is given, as well as a short explanation of how these types are currently being collected.
Link	http://www.oclc.org/research/activities/hiddencollections/borndigital.pdf

¹ **Manzuch, Zinaida.** *An analysis of the state-of-the-art in measuring the progress of digitisation of cultural materials.* pp33-35. See: http://www.numeric.ws/uploaded_files/NUMERIC_Desktop_Research_on_Digitisation_Studies26102007114936.pdf

3.4 Guidelines for web statistics

At the time of the NUMERIC project web statistics was still a very new area of interest for many institutions. In many ways it still is today. A recent study from Culture24 has found that this still complicates statistical comparison between institutions.² In order to achieve a certain degree of universality in the data concerning web statistics a set of guidelines would be in place.

Below, some reports, organisations, tools and guidelines are given that are intended to offer some help in setting up a strategy for web statistics and aid in interpreting them.

Title	<i>Let's Get Real: How to Evaluate Online Success?</i>
Developing or managing organisation	Culture24
Available since	2011
Background	The Report from the Culture24 Action Research Project. In it, insights are given into the ways in which institutions are trying to evaluate their online success.
Description	In this report basic steps for setting up a Google Analytics account can be found, as well as "10 key things to do," three of which are especially relevant to the present surveys.
Link	http://www.keepandshare.com/doc/3148918/culture24-howtoevaluateonlinesuccess-2-pdf-september-19-2011-11-15-am-2-5-meg?da=y

Title	[No title]
Developing or managing organisation	Web Analytics Association (WAA)
Available since	2004 (?)
Background	[no information]
Description	The Web Analytics Association offers its members (but also non-members) many tools to enhance their knowledge about web statistics. One of the tools that best suit the need of the ENUMERATE project is the list of definitions WAA published 2007: (http://www.webanalyticsassociation.org/resource/resmgr/PDF_standards/WebAnalyticsDefinitionsVol1.pdf)
Link	http://www.webanalyticsassociation.org

² Jane Finnis, Sebastian Chan, Rachel Clements. *Let's Get Real: How to Evaluate Online Success?*
<http://www.keepandshare.com/doc/3148918/culture24-howtoevaluateonlinesuccess-2-pdf-september-19-2011-11-15-am-2-5-meg?da=y>

Title	<i>Tien tips voor goede webstatistieken</i> [Ten tips for good web statistics]
Developing or managing organisation	Digitaal Erfgoed Nederland (DEN)
Available since	2009
Background	These tips are based on the report “More Digital Facts,” which, among other things, looked at web statistics.
Description	The English summary of the reports on which the tips are based can be found here: http://www.den.nl/getasset.aspx?id=Rapporten/Webstats_summary_ENG.pdf&assettype=attachments
Link	http://www.den.nl/bericht/2239

4 Conclusions

The overview presented in this deliverable is just a sample of the tools available. In constructing this overview an effort was made to include at least one of each tools per type of institutions (limited to archives, libraries and museums) and tools that can be used by all institutions. A further effort was made to make sure the tools are developed by authoritative organisations and people.

This overview is a work in progress. It will be added to as soon as new (or old) tools become available. New additions to this list will be added to the ENUMERATE Delicious account and supplied with the tags ‘harmonisation’ and ‘D2.4’

The ENUMERATE Delicious page can be found through the following link:

<http://delicious.com/enumeratesources>