

Project “ETQIL”

course

Digitizing Projects

Work book

Mats Dahlström, University of Borås

2014



Table of Contents

Introduction _____	4
Focus _____	4
Course design and previous knowledge _____	6
Learning outcomes _____	6
Course overview _____	7
Modules _____	8
Module 1: Digitization, purposes and strategies _____	8
Structure and tasks. Exercise. _____	8
Educational content _____	10
Module 2: Technology: text capture and encoding _____	12
Structure and tasks _____	15
Minor assignment: OCR _____	16
Educational content _____	19
Module 3: Use and assessment _____	20
Structure and tasks. Exercise. _____	21
Educational content _____	22
Major assignment: Digitization Plan _____	23
Form and size _____	24
Contents _____	25
Assessment _____	26
Course assessment _____	27
Instructions for trainers _____	27
Module 1 _____	30
Module 2 _____	31

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



Seminar _____	32
Module 3 _____	33
Major assignment: Digitization Plan _____	33
Seminar _____	33
Assessment _____	34
How to update the material _____	35
Preparations for trainers _____	35
Instructions for course designers _____	36

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



Introduction

This training course deals with cultural heritage digitization, specifically in the form of projects performed by libraries digitizing material with textual content. It will provide course attendants with a basic overview of motives and strategies, point to some examples of finished or ongoing digitization projects of national or international importance, take a small peek into some of the text-managing technologies behind them, and discuss the important issues of use, re-use and assessment of the digitized material, also from the point of view of research.

Since the emergence of large digital libraries and collections in the mid-90s and onwards, and not least in parallel to large private and public endeavours such as Google Book Search, Europeana and the Internet Archive, issues of quality and (re)usability of digitized text collections have been - and continue to be - high on the agenda of European cultural policies and research funding bodies. The need for competence development and the raise of awareness of the problems and opportunities surrounding cultural heritage digitization is significant.

At the same time, cultural heritage digitization is an area where many different interests and skills meet and can be applied in combination, such as technology, history, culture and sociology. It further provides an opportunity - even necessity - to strike a balance between, on the one hand, theoretical concepts and abstract principles, and on the other, hands-on skills and daily practices.

Focus

The course focuses digitization efforts at a national and international level and so, for the time being at least, turns a blind eye to the increasingly important and interesting issue of digitization at regional or local libraries or of private collections. Most attention is devoted to publicly funded libraries, such as national libraries, university libraries and to some extent other research libraries.

Further, since image editing and management are issues dealt with elsewhere in the course package, this course has been designed to focus on the digitization of

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



textual content and hence on textual document collections. Much more complex issues of technology and practices emerge if we would turn our attention to the digitization of sound or moving images, but this is beyond the scope of this particular course.

As for the course's module on technical implementation, this has to be necessarily brief and only provide a couple of examples of technology in projects digitizing textual content. It should be noted, though, that elements of hands-on technology, basic or quite cutting-edge, are present throughout the whole digitization "chain", and will differ considerably according to the specific purpose, material, level and needs of each individual digitization project.

The course is divided into three modules (with two exercises and one minor assignment task) and one overall, major assignment task.

The first module provides an overview of the digitization process, often conceived of as a chain of procedures, where the actual capture and editing of content are only two of many links. We look at some commonly referred motives for digitizing cultural heritage material, and some dominating strategies for carrying out the digitization.

The second module attempts a very brief look into two of the many technological phases in digitization, namely text capture through OCR and text encoding using XML and TEI.

In the third module, we turn our attention to what might happen once the material has been digitized - who is benefitted by the projects, who are the presumed users, are the projects used and do they open up for re-use or not? How can the general public be engaged in the projects and products?

In the major assignment task, you as course participant required to produce a small Digitization Plan for a particular library collection of textual documents. This Plan should preferably be written throughout the whole course and in parallel,

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



therefore, to the three modules. The Digitization Plan is presented and discussed at a seminar.

Course design and previous knowledge

The course is designed to take 40 hours of work, given that the participant already has some basic familiarity with cultural heritage digitization and digitized collections. Such knowledge includes an understanding of the many-sided task of publicly funded libraries and its role in contemporary society, its interest in cooperating with scholarly communities while also serving its patrons and the general public, its devotion to making the heritage available for contemporary and future generations while also adhering to its task of preservation and conservation of materials. As for the understanding of practical and technological issues, the course attendant should have a fair amount of computer literacy, some basic familiarity with metadata and the web as multifarious publishing platform. To really get the most out of module 2, the attendant will benefit from prior knowledge of text encoding, XML and character encoding, although this is not a formal requirement.

The course consists of a mix of individual readings, exercises, assignments to be performed and demonstrated, and group seminar discussions. The course does not offer much room for extended literature reading and the study of many different projects, so the attendants are strongly encouraged to share further tips and pointers with each other on interesting texts to read and projects to study and be inspired by.

NB! Even though all the external educational content (and most of the recommended additional material, too) is available online free of charge in some version, most of it is covered by copyright restrictions.

Learning outcomes

After completing the course, the participant will be able to

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



Learning Outcomes	
Knowledge	Skills
... critically analyse specific motives and strategies for digitizing textual documents in library collections	
Identify common motives for cultural heritage digitization and some general strategies for performing it, on a national and international library level, while also critically reflecting on those motives	Implement critical knowledge of motives in a concrete digitization plan for a particular collection of textual material, according to a strategy that meets the overall purpose of the project and the characteristics of the material
...display an understanding of some current technologies for text capture and text encoding and their consequences for successfully performing a sustainable digitization project of textual material	
Understand the basic procedures of text capture and text encoding in general and of OCR and XML/TEI in particular. Account for some fundamental possibilities and problems of these technologies with respect to different kinds of source material and sustainability.	Perform an experimental but simple evaluation assessment of current OCR software and analyse non-trivial consequences and problems with their varying performance rates. Declare proper strategies and levels of text encoding for a specified source material.
... understand the interdependence between a digitization project's design and the needs of particular user groups for the digitized collection, and to identify various ways of assessing the value of digitized collections for users and research	
Understand some important ways in which digitized text collections serve the needs of its intended user groups, such as the general public or particular research communities. Identify possible ways to assess the value of a digitization project.	Implement this knowledge in a concrete digitization plan for a particular collection of textual material, according to a strategy that meets the needs of specified user groups and stating the use, re-use and research potentials of the digitized material.

Course overview

Modules	Activity	Time

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



Module 1: Digitization, purposes and strategies	Reading/viewing educational content Exercise	3.5 h 2 h
Module 2: Technology	Reading/viewing educational content Minor assignment: OCR Seminar	4.5 h 6.5 h 2 h
Module 3: Use and assessment	Reading/viewing educational content Exercise	3.5 h 2 h
Major assignment: Digitization Plan	Digitization Plan Seminar	13 h 3 h
<i>Total</i>		<i>40 h</i>

Modules

Module 1: Digitization, purposes and strategies

The first module takes an overview of the digitization process, often conceived of as a "chain" of procedures, where the actual capture and editing of content are only two of many links. We look at some commonly referred motives for digitizing cultural heritage material, and some general strategies for carrying out the digitization. Focus is on cultural heritage digitization (primarily of textual documents) in memory institutions and cultural sectors.

Structure and tasks. Exercise.

Start by listening to the lecture (Dahlström 2014). Then read Coyle for her take on the strategy of mass digitization, as exemplified with i.a. Google Book Search. Make sure you understand the distinction between text-based digitization and image-based digitization, conceived of as general *methods* for digitization projects, and between mass digitization and critical digitization. Familiarize yourself with the pros and cons of both large-scale and small-scale digitization. Consider the

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



different avenues that different digitization projects can take in terms of scale, level and quality, depending on why the project has been launched to begin with, what its motives are, who runs it and for what purposes.

When listening and reading, keep in mind questions such as: Why do libraries in general engage in digitization projects of their holdings in the first place? Think about the state of digitization projects in your own country, and perhaps at your own work place. How are they financed and why? Is it possible to combine different purposes within one and the same project? Which purposes and strategies might be difficult or even impossible to combine? Does quality and quantity exclude one another in digitization projects? If so, why? If not, why not?

Then choose two digitization projects at national libraries or research libraries in your home country, and study them critically with the questions below in mind. The projects should preferably have made available its digitized material primarily openly on the web – i.e., do not choose a project whose digitized material is primarily reserved for “in-house” use within the institution, or locked behind passwords or payment mechanisms, or limited to being accessed only if and when a user visits the institution physically onsite.

See if you can find specified declarations of purpose for the projects, either directly through an available policy document, or by asking people connected to the project, or by reading a report or evaluation of the project, if there is one. (If you cannot find any such documentation - what does that say about the institution and the project, and why do you think that is?). You should also study the digitized collection itself and draw conclusions about what possible *implicit* motives and strategies you can discern for the digitization project.

Questions to consider:

- What seems to be the main motive for the digitization project – why was it launched in the first place? How does the institution “sell” the project, its purposes and its use? Does for instance preservation play a significant part,

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



or is it rather to make the material available and distributed that drives the project? Can you think of other reasons for this project?

- Are the motives of the project consistent with the way that digitization has been performed and material been made available?
- Is the project part of (or in cooperation with) another project or activity with similar character or agenda, within or outside of the institution?
- Has the project been able to re-use digitized stuff from earlier/other projects?
- Do you think the project affects the availability and use of the physical source documents for patrons?
- Which user group appears to be the main patron for the project? Has the digitization itself been performed at a relatively low or high technical level? Is it your conclusion that the project's level of ambition and technical standard matches that of the declared user group for the project?
- Is the download of particular dedicated software needed to access the digitized material? Are parts of it protected behind password or payment mechanisms? Is the collection easy to navigate? Is this level of availability in harmony with the specified purpose and intended user groups?

This exercise and the considerations are primarily for your own benefit, and are not directly examined in a special task. The module contents are however examined *indirectly* in the Digitization Plan, where you need to devote a particular section to describe (and preferably argue for, if you see the need) the purposes and strategies chosen for your described digitization project.

Educational content

Coyle, K (2006). Mass Digitization of Books. *Journal of Academic Librarianship*, 32(6): 641–645.

Dahlström, Mats (2014). *Introduction to digitization*. (Lecture recording)

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



Recommended literature & useful sites:

The recorded lecture (Dahlström 2014) included in this module is necessarily brief. If you would like a more expanded text version, you can read:

- Dahlström, Mats (2011). Editing Libraries. In: C. Fritze, F. Fischer, P. Sahle & M. Rehbein (Hrsgg.), *Bibliothek und Wissenschaft. Vol. 44: Digitale Edition und Forschungsbibliothek*. Harrassowitz. PP. 91-106. (A post-print is available for free at: <http://bada.hb.se/bitstream/2320/9764/2/buw-44-dahlstrom.pdf>)

For those of you interested in a thorough text book on cultural heritage digitization, there are many to recommend. A useful book is:

- Terras, Melissa (2008). *Digital Images for the Information Professional*. London: Ashgate.

This has become a much cited work on digitization within memory institutions and cultural heritage digitization. It covers most of the basics but also proceeds to discussions and analysis, providing many examples. A newer book is:

- Bülow, Anna & Jess Ahmon (2011). *Preparing Collections for Digitization*. London: Facet.

However, Bülow & Ahmon does not cover the topic of metadata for digitized material. Other than that, this book covers about the same topics as Terras' but with more emphasis on preservation and archival issues.

Both these printed books are relatively expensive to acquire if you cannot get hold of them through a library loan, and there are of course several good manuals and guidelines for free on the web - although these are much more hands-on and often tailored to the needs of specific communities. Some more general guidelines are:

- Federal Agencies Digitization Guidelines Initiative (2011).
<http://www.digitizationguidelines.gov/guidelines/digitize-planning.html>

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



- DFG (Deutsches Forschungsgemeinschaft) (2009): Practical Guidelines on Digitisation
http://www.dfg.de/download/pdf/foerderung/programme/lis/praxisregelIn_digitalisierung_en.pdf
- Minerva EC (2006-). <http://www.minervaeurope.org>

See especially Minerva's *Good Practice Handbook*, and their technical guidelines. These are handbooks for most steps in the digitization process, composed primarily with European memory institutions in mind, and provide several useful and practical suggestions.

- National Library of Australia (2010). *Collection Digitisation Overview*.
(available at: <http://www.nla.gov.au/digital/>)

This guide overviews all the different steps in a digitization project. Please note that the web page at the URL is a start page, and that there are following pages on e.g. policies, priorities, preservation and workflows.

Module 2: Technology: text capture and encoding

The content concerning text capture and OCR in this module is largely based on material developed by David Hansson (Systems Analyst at Karolinska Institutet, Sweden) and Gunilla Wiberg (IT Coordinator at Lund University Library, Sweden) for a larger (15 ECTS) digitization course at master's level at the Swedish School of Library and Information Science, Borås.

Technological competence is crucial throughout the whole digitization process and is implemented in e.g. collection surveys, equipment acquirement and adaption, image capture and editing, text capture and editing, text encoding, metadata, publishing and marketing, crowdsourcing, log analysis, and preservation and maintenance tasks. Which specific technology skills are needed will differ with each individual project, its purpose, material, level and needs. In this course, there is only time for a quick peek into the phases of text capture and encoding. The

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



previous module introduced the distinction between image-based digitization and text-based digitization as two major strategies, and so the latter one is being exemplified in this module.

The phases of text capture and text encoding are connected to the important aspects of interoperability, standards and sustainability - that is, making sure the digitized collection as data can be maintained and exchanged across time and contexts, synchronized and perhaps merged with other similar collections, and be subjected to e.g. indexing, search and extraction/transformation applications that are based on a predefined set of formats and standards. Often, sustainability issues concerns metadata formats, but we will use this occasion to highlight the particular necessity of working with standards and interoperable formats on the level of character representation and full text encoding (next module) as well. This issue has previously often been neglected in many national digitization projects, which settle for scanning text pages as images only and putting these images on the web, and where the user cannot even search the textual content of the images because text capture and editing has not been implemented in the project.

It is important to understand the delicate balance between input and output in digitization projects. The efforts and the quality the you invest in the input and editing phase are crucial to what use and benefit the material will have for various user groups, and so this module on technology not only reflects on the purposes and strategies mentioned in the previous module but also points to the value of collections discussed in the next module on use.

To begin with then, there is the issue of the transition from digital image to digitized text; once you have captured contents in a source document through e.g. scanning, ending up with a digital image, how do you proceed to turn this image into a machine-readable text that you can search and encode? This involves various methods for text capture, particularly (but not exclusively) through the means of Optical Character Recognition (OCR) and character encoding.

Character encoding is a system for computers to store and interpret symbols. Each symbol is represented by a number. A symbol might be a letter (such as “a”)

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



or a trademark sign (™). Different character sets have historically caused problems between different computers and human languages, because they are limited to 256 characters. Hebrew character sets were not the same as western European character sets. Mac and Windows computers also differed in their characters sets. Unicode is trying to solve some of these problems and has a total of 65 000 different characters.

OCR is a process that converts a (scanned) digital image into digital text. The image is analyzed by the computer. Paragraphs, text lines, words and letters are identified in the image. Quality issues affect the result: There might be problems with bleed-through in the print in the image. The image might also be lighter and too white in some areas, or too dark in other areas. Fuzzy matching may be used to analyze patterns in words and decide whether a symbol should be a letter or digit. Lexicons (as modules within the OCR software itself) are used to spell check the output and identify words.

OCR might be used to convert images from books and papers into searchable and editable texts. OCR interpreting older material is not of the same commercial interest as OCR interpreting newer material. A large EU project called *Impact* has been aimed at improving the OCR interpretation of historical material (see the list of reference literature).

Once you have a machine-readable text transcription, this usually needs to be proof-read and corrected before being subjected to text encoding. Text encoding is introduced both in the lecture (Dahlström 2014) and in the course literature. It entails assigning specific codes to a transcription text in a standardized and transparent manner, opening up for the texts to be used and re-used, exchanged, extended, extracted, merged, classified and transformed.

After a long history of text encoding, the emergent standard for markup has become XML (eXtensible Markup Language). There is no guarantee XML will remain the standard for all future to come, or even remain at all, but it is currently the best we have, and the closest we come to an encoding standard framework.

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



XML comes in many different forms and applications for specific communities and types of material. For humanistic and historical material at least, a particular XML-based application has become a de facto standard: TEI (Text Encoding Initiative), currently in its fifth version, P5. TEI is governed by a large international joint effort, the TEI Consortium (TEI-C) - and on its website (<http://www.tei-c.org>) one can find all the necessary and updated specifications and guidelines.

Structure and tasks

Start by reading the general introduction by Deegan and Tanner (2004) to the conversion of printed materials and to various methods for text capture. This is followed by the overview of OCR by Tanner (2004). These two texts are written by a couple of Britain's foremost experts on digitization. They present the basic problems and solutions, with several good examples on OCR.

Continue by listening to the lecture (Dahlström 2014) introducing text encoding and XML. The lecture provides an overview of some general strategies for text encoding and subsequently looks at XML, which has become the dominant standard markup format for encoding digitized textual content. (For those of you wanting to acquaint yourself a bit further with how XML works, take the XML tutorial at <http://www.w3schools.com> or see the text book by Ray, 2003).

As for TEI per se, there are clear specifications and (quite bulky) guidelines on what, why and how to encode. These are maintained by TEI-C as mentioned earlier. There is however not yet a specific standard text book on TEI, but there is at least one very good introduction and manual on the web: *TEI by Example*. Leading experts on TEI have brought together this teaching resource. They write in an understandable manner about the basics of TEI. For this course, you are required to read sections 1-5 in "Module 0: Introduction", which should provide you with a general understanding of how TEI works and what it can offer. For those interested, *TEI by Example* also includes further and more specified sections on TEI along with quizzes, exercises, and loads of good examples. So this is really a beginner's "must" guide to TEI. Those wanting to know more about where TEI came from can study Vanhoutte (2004).

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



The purpose with this course content on text encoding is not to prepare you to perform actual hands-on TEI encoding yourself, but rather to understand *why* texts are encoded with TEI and what added value TEI can bring to the digitized material. Having studied the content on text encoding in general, then, you should be able to understand at least the background, rationale, purpose, and basics of text encoding, primarily from the perspective of international standards and with an eye to written cultural heritage materials. Having studied the content on XML and TEI in particular, you should be able to understand the basic possibilities and benefits of those technologies for digitization projects of textual material, as well as acknowledge some limitations and possible risks of using them for various institutions, projects and material types.

Further, this knowledge will of course prove valuable for you when writing the sections in the Digitization Plan, where you are required to devote a section on text capture and encoding issues.

Minor assignment: OCR

To get a taste of the possibilities and problems of applied technology, the module contains a small experimental assignment on OCR. This can be performed either as individual or group-based exercises (if for instance you attend the course together with colleagues from your workplace). The purpose of the assignment is not to point to "correct" answers, instead the results, problems and observations are to be discussed at a group seminar or webinar moderated by the trainer.

For this assignment, you are to transform a scanned image of a book page into machine-readable text by running Optical Character Recognition (OCR) software, and then attempt to analyse the various results of the process. Document your results and in write down the questions and problems you came across during the exercise. You are not required to report or submit the empirical results to the trainer, but you are to bring them and the questions and problems you came across to a seminar, where these should be discussed (see further below). Perceived failures are equally important as perceived successes.

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



You need a source document to work with. For your convenience, the module provides already prepared image scans for you. They are images representing a single page in a scanned book (mostly in Gothic type). It has been processed in black/white and in colour, and in low resolution (100 dpi) and in high resolution (600 dpi). Different software requires different file format input (usually JPG or PDF). Therefore, the files are available as both PDF and JPG and you are to pick the format that fits the software you intend to work with (or both, if that is possible). (Should you prefer another (similar) source document to work with, you need to acquire it and scan it yourself.)

You are now to turn these scans into machine-readable text by processing the files in at least two different OCR programmes. Run all scan versions. There are some free OCR alternatives. One example is FreeOCR (Windows): <http://www.paperfile.net/>. Two examples of an online service are <http://www.onlineocr.net/> or <http://www.free-ocr.com/>. There are also of course commercial software brands, Adobe Acrobat Professional is one. A much used piece of software is the commercial Abbyy FineReader, which is available in a 15 day trial version at <http://finereader.abbyy.com/trial/> (Windows) and http://www.abbyy.com/finereader_for_mac/trial/

When you have run the scans in the software, you receive an output with the resulting text. Output forms differ between the various programmes. Some online OCR services produce a simple downloadable text file, whereas e.g. Adobe Acrobat Professional provides a PDF output.

In this assignment, you are likely to come across technical problems. For instance, not all of the different scan versions may be processable with one and the same software (e.g. because of too low or too high resolution in the image file, or because of a rotated scanned page). Make a note of such problems. If you want to attempt to solve the problems, you might need to try more than one piece of software. The Abbyy FineReader is perhaps the most accomplished of the programs mentioned above. Don't be afraid to test and experiment with different software. Study the help sections for the software you are using and learn what is doable and what is not. Also, be aware that OCR processing may take time,

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



particularly when using free online services. For some services, you might need to register and then wait for confirmation.

Feel free to experiment with the OCR settings, depending on which program or service you use. Examples might be language, or downsizing the resolution (if using Adobe Acrobat Professional). Try the scans with different settings until you are fairly satisfied with the results. Check the error rate, note how the results and performance vary with different scan versions and different software, make observations, and draw conclusions both about the use of OCR for older material and the problems you experienced.

Seminar

Attend the seminar arranged by the trainer and present the observations you found most interesting, focussing on the problems you encountered, how they might be explained based on what you know about OCR, and if and how you tried to solve them.

Some discussion topics may be the following:

- Does the result differ between different OCR engines? If so, how? Why, do you think?
- Did the pre-set resolution affect the result? In what ways? Why?
- Did the differences between colour and black/white affect the result? In what ways? Why?
- If you experienced a difference between the OCR of Gothic type and that of Roman letters in the same document, what might explain these differences? Which one is best interpreted by the OCR engine? Why?
- How do different languages and accents (such as é and ê) affect the OCR result?

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



- Does show-through in the print affect the OCR result?

Assessment

To pass this module, the participants are expected to present at least three observations and problems each from the assignment at the seminar as well as to engage in the discussion of the other attendants' observations and problems.

Educational content

Dahlström, Mats (2014). *Introduction to text encoding and XML*. (recorded lecture)

Deegan, Marilyn and Simon Tanner (2004). Conversion of Primary Sources (ch. 32). In: Schreibman, Susan, Siemens, Ray & Unsworth, John, eds. *A Companion to digital humanities*. Oxford: Blackwell. PP. 488-504. (available at: <http://www.digitalhumanities.org/companion/>)

Tanner, Simon (2004). *Deciding whether Optical Character Recognition is feasible*. London: King's College. 11 pp. (available at: http://www.odl.ox.ac.uk/papers/OCRFeasibility_final.pdf)

van Branden, Ron, Melissa Terras & Edward Vanhoutte. *TEI by Example*. <<http://www.teibyexample.org>> (2014-03-15)

Reference literature & useful sites:

Burnard, Lou, O'Brien O'Keefe, Katherine & Unsworth, John, eds. (2006). *Electronic textual editing*. New York: MLA/TEI. (available at: http://www.tei-c.org/About/Archive_new/EET/Preview/). Chapters:

- Eileen Gifford Fenton and Hoyt N. Duggan: Effective Methods of Producing Machine-Readable Text from Manuscript and Print Sources
- John Lavagnino: When not to use TEI

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



- Sebastian Rahtz: Storage, Retrieval, and Rendering

The IMPACT project: <http://www.impact-project.eu/>

- On OCR problems for historical material, see IMPACT's project description (particularly the first 7 pages) at: http://www.impact-project.eu/uploads/media/IMPACT_Description_of_Work_public_version.pdf
- A good and concrete presentation, with lots of illustrative examples: http://www.impact-project.eu/uploads/media/Apostolos_Antonacopoulos_Digital_Restoration.pdf
- A good presentation on word lists: http://www.impact-project.eu/uploads/media/Katrien_Depuydt_Historical_Lexicon_Building.pdf

Ray, Eric T. (2003). *Learning XML*. 2. ed. Sebastopol: O'Reilly.

The Text Encoding Initiative Consortium: <http://www.tei-c.org>

Vanhoutte, Edward (2004). An introduction to the TEI and the TEI consortium. *Literary & Linguistic Computing*, 19(1): 9-16.

w3schools. *XML tutorial*. <<http://www.w3schools.com>>

Module 3: Use and assessment

In the third module, we turn our attention to what might happen once the material has been digitized - who is benefitted by the projects, who are the presumed users, are the projects used and do they open up for re-use or not? How can the general public engaged in the projects and products? How can the projects and the digitized collections be assessed?

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



In the module, you are to look back on what has been done in the digitization process, its outcomes, who can use the collections for what end, and how all this relates to the original intentions, scopes, audiences and reasons for digitization in the first place. Thus, the module ties back to the motives and strategies in module 1 and the technology issues in module 2.

Structure and tasks. Exercise.

Listen to the lecture by Dahlström (2014). Then read Seaman (2003) and Palmer (2004).

Then return to the two digitization projects you chose for module 1, and study them critically with the questions below in mind.

Consider the potential these projects might have for the *general public*:

- who gets to do what with the collections and for what purpose?
- how can the collections be an instrument for engaging the general public and fulfilling the societal role of the library?
- how can these libraries work to increase both the usability and the actual use of their digitized collections?
- using what technologies and to what extent do the digitized collections allow for incorporating user-generated content or other forms of user interactivity, if at all?
- using what technologies and to what extent are they made available as "open source", if at all?
- have the projects been evaluated, and is there any available documentation (such as a project report) of this? If so, what were the conclusions and recommendations for future work?
- has any effort been made to measure user statistics, for instance through a log analysis? If so, has a report of that analysis been made available? Are the statistics data themselves available? Are these types of statistical data and log analyses common or rare within the cultural heritage sector of your country, would you say? Why?

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



Consider the potential these projects might have for *research*:

- to what extent are the collections based on research and to what extent can library digitization in general work with and support scholarship and research in the humanities or the social sciences? Are subject scholars (such as philologists, historians, digital humanists, literary scholars, book historians) involved in the digitization process?
- can the collections be of significant value for research? What is required for them even to be regarded as research data in themselves?
- can library digitization such as this lead to new research and what would be needed to support such a development?
- are the digitized collections accessible and usable to researchers, would you say? And are there any indications they are actually being accessed and used by those communities? What do you think is the general state of affairs in this respect in your country? Why?

These studies and considerations are primarily for your own benefit, and are not directly examined in a special task. The module contents are however examined *indirectly*, in the Digitization Plan, where you need to devote a particular section to the issues of the intended users and types of usage and benefits of your described digitization project.

Educational content

Dahlström, Mats (2014). *Digitization - access, use and research*. (Recorded lecture).

Palmer, Carole (2004). Thematic Research Collections. In: Schreibman, Susan, Siemens, Ray & Unsworth, John, eds. *A Companion to digital humanities*. Ch. 24. Oxford: Blackwell. (available at: <http://www.digitalhumanities.org/companion/>)

Seaman, David (2003). Deep Sharing: A Case for the Federated Digital Library. *EDUCAUSE Review* 38.4: 10-11. (available at: <http://net.educause.edu/ir/library/pdf/ERM0348.pdf>)

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



Reference literature & useful sites:

Tanner, Simon (2012). *Measuring the Impact of Digital Resources: The Balanced Value Impact Model*. King's College London. Available at:
www.kdcs.kcl.ac.uk/innovation/impact.html

Warwick, Claire et al. (2006). *The LAIRAH Project: Log Analysis of Digital Resources in the Arts and Humanities. Final Report to the Arts and Humanities Research Council*. London: School of Library, Archive and Information Studies, University College London. Available at:
<http://www.ucl.ac.uk/infostudies/clairewarwick/publications/LAIRAHreport.pdf>

Major assignment: Digitization Plan

The main assignment in the course is to design a project Plan for digitizing a specific source material with textual content. This material can range from a single document to a large collection in need of digitization, and the material can be housed by a particular library, e.g. at your workplace, or be a private collection (although the former is recommended). If you do not have a particular real collection in mind or access to it, it is possible to design the Plan for a fictive document or collection of documents.

As for selection of material and strategy, you can either focus on volume or on depth:

1. If you focus on *volume*, your Plan might e.g. concern a digitization of a large collection of objects, comprising several thousands of objects with some degree of internal variation with respect to e.g. size, type of binding, and condition of items.
2. If you focus on *depth*, your Plan might concern a "critical", high-quality digitization of a single or a few complex object(s), such as manuscripts or composite documents.

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



It is your responsibility to present trustworthy arguments of why the selected source collection needs the particular methods, process and project design that you argue for in your Plan.

To support you in writing the Plan, you might want to consult, firstly, Bülow & Ahmon (2011) or Terras (2008) or a similar overview of digitization processes, secondly, the guidelines and manuals referred to as suggestions at the end of module 1, and thirdly, the required and the recommended literature in the course modules as well as any additional literature you deem helpful for your particular project.

Form and size

The form and size of the Plan can vary considerably, depending on your interests and ambitions as well as on the characteristics of the source material, as long as the material to be digitized consists of textual content (books, manuscripts, letters, newspapers, journals, etc). The Plan can take on the form of e.g. a regular essay in a word processing format (such as MS Word or Open Office) of a few pages, or a spread sheet accompanied with explanatory text, or a slide presentation, or a presentation on the web; choose the method of delivery that you feel most comfortable with and would use if you were to propose a digitization project in front of your library's board of directors (or equivalent). Remember that the total amount of hours you are expected to spend on the Plan is ca 13, so do not overdo your work. Still, make sure the Plan contains what is required (see below) and that it is realistic and balanced with respect both to the material that is to be digitized and to the library (or other institution) that is to perform the proposed digitization. Thus, you need to base the Plan in the conditions, competences, resources and needs of a specific library. Do not write a utopian “dream plan” - any digitization project needs to compromise between what it would like to do (given limitless resources) and what the available resources allow for.

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



Contents

The Plan must reflect the three modules in the course. This means that at the minimum, it must consist of:

- a short description of the source material (total number of objects, media type(s), type of content, size, current placement, ownership, condition, formats, etc)
- a declaration of why this (selection of) material needs to be digitized, what the main purpose(s) of the project is
- a declaration of who or which institution is to perform the digitization
- a brief presentation of the strategy chosen: do you recommend an ambitious, high-quality, manual digitization process, or a quick-and-dirty machine-driven mass digitization strategy, or something in between?
- a brief walk-through of the proposed digitization process and its steps
- a rough estimation of what resources the project needs in order to be realized, with respect to e.g. time, costs, human resources/staff, hardware (computers, cameras, scanners, cradles etc) and software, collection survey, licenses, rooms/studios/labs, competences and skills, clearing of copyright issues, metadata, etc. This does not need to be an extensive section, the important aspect here is for your Plan to come off as being realistic and doable, given the institution you have declared as performing the digitization
- a particular section on the planned text capture and text encoding phases within the project. Try to identify any particular problems the project is likely to come across due to the characteristics of the source material.
- a declaration of who the intended users are, why they need this particular material to be digitized, and how the project, the selection and the strategies and methods have been chosen to best suit the needs of this particular user group.

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



- an identification of possible risks during the project, identifying factors that might prove to have negative effect on the results of the digitization process.

Assessment

The Plan (or a link to the Plan) is submitted/made available through the platform to the other course participants. This is followed by a seminar, where your Digitization Plan is briefly presented by you and then discussed by the seminar. The trainer follows up by specifying strengths and weaknesses of the Plan and notifying whether or not it has passed. This is based on whether or not the Plan:

- is designed according to the instructions, and has at least the contents as specified above
- is comprehensibly and understandably presented
- displays realism and relevance with respect to the material chosen, the proposed motive for the project, the available resources, and the needs of the targeted user group(s)

The Plan deserves *special merit* if:

- the criteria have been met on a high and critical level
- it has a reflective and problematizing approach, where you display a particular ability to identify possible technical or administrative risks and potential problems for the project to come true, and where you are able to propose good solutions to such problems.
- it contains concrete ideas about how the digitized collection could be evaluated, and its use measured
- there is a reflection on the long-term benefits of the proposed project - for users, for the institution, for the material itself, for other similar projects, etc

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



Course assessment

The course contains two assignments (one minor on OCR and one major as a Digitization Plan), two exercises and two seminars. The minor assignment is not to be submitted, but will be discussed at one of the seminars. The major assignment is to be submitted and then discussed at the second of the seminars. The two exercises are not examined and assessed per se, but are to provide impetus for the respective sections in the Digitization Plan.

The successful participant is expected to have concluded the following tasks in a satisfactory manner:

- Fulfillment of the minor assignment, as demonstrated at the seminar, where the participant presents at least three observations and problems from the assignment and engages in the discussion of the other participants' observations and problems.
- Fulfillment of the major assignment, by submitting a Digitization Plan that meets the requirements and by presenting and discussing the Plan at a seminar.
- Active participation in both seminars. 'Active' is defined as having contributed to the discussion and/or made presentations.

Instructions for trainers

An educational idea behind the design of the course is to combine individual readings, case studies (the exercises) and an experimental assignment (OCR) in order to help the course participant make practical use of all this impetus in the form of an overall and concrete Digitization Plan, which hopefully will be of use not only to him or her individually, but also to the institution where he or she works. Ideally, the Digitization Plan can be further used and referred to in other courses as well, for instance M1, M4 and M7.

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



Both the assignments and both the exercises are designed as individual tasks. However, if you (and the participants) see fit, they can equally well be performed as group tasks. If you decide for this, state this already at (or before) the start of the course, and declare to the participants that each group member's contribution in the group assignments must be possible to identify clearly and individually in order to be assessed.

The most important learning outcome for the participants is to understand the delicate balance between the quality of the input and what quality one will expect from the output. This should become evident in their Digitization Plans. You should not assess these Plans in the sense that a high-quality project necessarily deserves better merit than a project that is less ambitious. A high-quality ambitious digitization project (and plan) is not necessarily "better" than a small-scale, quick-and-dirty project with a low degree of technical implementation or manual hands-on labour. It depends on the needs of the material, the digitizing institution, and the presumed user groups. Some users might e.g. only be interested in having access to roughly scanned images of the material to be able to decide whether or not they will be able to find what they are after by travelling to the library housing the source documents, other user groups will need to have fully TEI encoded and proof-read text transcriptions with high resolution image scans side by side. Some projects will be more experimental in nature, testing a particular technology, software or method, and where e.g. long-term maintenance and archiving of the digitized material might be of less importance. The participant should demonstrate in the Digitization Plan how the proposed project aims to manage this balance between input, available resources, and expected output and value.

As for the technical parts in module 2, the primary aim should not be that the participants must demonstrate a high degree of technical skill, that is not the purpose. Module 2 and its assignment is only meant as an example of how technology can be implemented, what benefits and risks are involved, and - again - how there is an intrinsic balance between on the one hand the efforts you put down in the input and editing phase and on the other hand the value and use you can expect from the digitized material. In the assignment, the ability to recognize

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



problems is more valuable than to achieve software performance "success". Also, there might be cases where OCR and text encoding are not very good options, due to e.g. the characteristics of the source material (such as hand-written letters or text that is illegible) or the resources and skills of the digitizing institution, and it is good if the discussion seminar picks up on such issues as well.

The major assignment, the Digitization Plan, is to be written *in parallel* to modules 1-3. This means the participants should get started on the Plan already early in the course, and the work is to be intensified towards the end of the course. Be very clear about this to the participants. A particular discussion thread on the ongoing work with the Plan is highly recommended and should be monitored throughout the course by you as trainer. This should not mean that you take on the role as "supervisor" for each Digitization Plan. Rather, encourage the participants to ask questions, make suggestions, and help *each other* in the Digitization Plan thread - each participant is likely to come across the same kinds of ideas and questions as other participants, and they are therefore in a good position to help each other.

Your role in the course is to interact with the participants in order to support and assess their learning. This includes the following:

- Encourage the participants to introduce themselves on the learning platform (unless everyone already knows each other) (c. 30 minutes)
- Monitor possible discussions on the learning platform and care to questions if there are any, either be answering yourself or by encouraging the participants to help each other (the group can decide on voluntary discussions) (time depends on amount of questions and discussions)
- Encourage the participants to declare on the learning platform what projects they have chosen to study in the two exercises (module 1 and 3) and what observations they have made (this is not required, but feedback from other participants on this will help them go through with the exercises and thus will support them in writing the Digitization Plan) (c. 30 minutes)
- Encourage the participants to declare on the learning platform what documents or collections they have chosen for their Digitization Plan, make sure there is a thread for this topic, and where they can pose questions, comments, and follow-up during their work with the Plan. (this is not

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



required, but feedback from other participants on this will help them produce the Digitization Plan) (c. 30 minutes)

- Arrange and moderate the seminar (or webinar) in module 2. Make sure each participant presents his or her observations as required and also contributes to the discussion of the other participants' presentations (c. 2 hours, excl. preparation time)
- Instruct the participants to submit their written major assignment, the Digitization Plan (or a link to it), through the learning platform at a given deadline in due time before the seminar. After submission, draw up a schedule for the seminar where each Plan is allotted equal time for presentation and discussion. If you see the need, the participants can be assigned special commentator tasks to each others' Plans. If so, you can specify this in the schedule. (c. 30 minutes)
- Arrange and moderate the seminar (or webinar) for the major assignment, the Digitization Plan. Make sure each participant briefly presents his or her observations as required and also contributes to the discussion of the other participants' presentations (c. 3 hours, excl. preparation time)
- Assess the course:
 - keep track of active participation in the seminars
 - assess the major assignment and provide brief comments to the author (c. 1 hour/Plan)

Module 1

Initiate the course by presenting yourself and inviting the participants to introduce themselves briefly on the learning platform.

The trainer's main involvement in this module is to monitor possible discussions and answer questions if there are any. Read the assigned course literature and listen to the recorded lecture. Familiarize yourself with some of the recommended literature, such as the guidelines and either Terras' book or the book by Bülow & Ahmon. Doing the exercise yourself (studying two cases of digitization projects) will be helpful in order to help the participants and to answer questions.

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



Module 2

At the beginning of module 2, make sure there is a learning platform discussion thread on both the module 2 and the minor assignment on OCR. The participants are likely to come across (and are indeed encouraged to focus on) several problems, ranging from small and trivial (how do I download this or that, can I use this book scan instead of the prepared scans, where can I get access to this or that software, etc), to tricky technical and processual problems (why is my output from software X in partly unreadable text, why doesn't this file format work in software Y, etc). They will likely need to ask and discuss these matters with each other on the platform, and it is only beneficial to all if they are able to initiate such discussions even before the seminar.

Go through the required reading and the recorded lecture. Familiarize yourself with the proposed OCR software, download the scans and perform the OCR procedure yourself to learn about the successes and failures the participants are likely to come across. You must certainly not aspire to attain an expert level on OCR and encoding, but are encouraged to explore this field at the level of the participants, which will help you engage in the seminar and assess their contributions: are their observations and questions "reasonable" and logical?

Decide beforehand on your planned schedule for the seminar and what is expected of the participants. Remind them that they do not need to submit anything before the seminar, but are only required to bring their observations of questions, problems and possible solutions to the seminar, present and discuss these at the seminar and engage in the discussion of the other participants' presentations. Also, be aware that although pre-prepared scans are provided, the work book instructions open up for the possibility for participants to work with their own book scans instead. If so, ask them to check this with you early on - the material they choose instead must offer some degree of difficulty similar to the Fraktur etc in the provided scans. OCR'ing scans of printed books in plain Roman text in clear resolution will probably offer the software little problems if at all, and so there will be little to discuss in the seminar.

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



Seminar

Begin by pointing to the proposed schedule, and then let the participants present their findings (and also what material they have worked with, in case they chose a different material than the prepared scans), and help them to keep the time agreed upon. Note active participation for assessment purposes. Let them know by the end of the seminar or after the seminar whether or not they have passed the assignment.

The instructions in the work book provide some questions that the participants can address during the seminar. There are even further discussions that might arise during the seminar:

- What happens with the OCR if an image is skewed (e.g. leaning, distorted)? Why?
- How does advanced layout on a page affect the OCR result? Why?
- How do old forms and variants of spelling match modern lexicons in the OCR process?
- What might be the obstacles to having OCR interpret material such as old card catalogues? How, do you think, would the OCR of notes compare to that of typewriting?

Prepare for the seminar by thinking about these questions and possible answers to them.

For the final 20 minutes or so of the seminar, initiate and moderate a general discussion on the possibilities and problems of text capture and OCR. Encourage the participants to draw conclusions about the findings during the seminar and the kinds of material and collections they are working with at the respective work places. Would OCR be a valid option for them and their material? Would the software they have tested be valid, or does it provide too many shortcomings? In what respects?

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



Module 3

As in module 1, the trainer's main involvement in this module is to monitor possible discussions and answer questions if there are any. Read the assigned course literature and listen to the recorded lecture.

Doing the exercise yourself (studying the same two cases of digitization projects as in module 1 but with an eye to use and assessment) will be helpful in order to help the participants and to answer questions.

Major assignment: Digitization Plan

Declare already at the start of the course that the Plan is to be written in parallel to the three modules. Make sure there is a discussion thread on the platform where the participants declare their chosen material and where they can post questions and comments and ask the help from each other.

If you see the need, specify the allowed formats and the maximum length of the Plan (the work book instructions are intentionally liberal on this point).

Seminar

For this seminar finalizing the course, the participants are to submit their Plans (through the learning platform) beforehand. Decide whether or not you as trainer will assess the Plans before, during or after the seminar. Also decide if you want the participants to have a look at some or all of the other participants' Plans before the seminar, thus paving the way for a peer commenting schedule. Depending on which, set a *submission deadline* that provides you and the participants with reasonable time to read the Plans before the seminar.

Decide on a suitable maximum length of the presentations. A suggested time length for the whole seminar in this work book is three hours, but this can vary depending on the number of participants, the format of the seminar, and the time available.

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



The seminar will primarily be a presentation of the participants' suggestion for a realistic (but still fictive, of course) Plan. Although this will be a suggestion and possibly a work in progress, their work should be as realistic as possible according to the available resources declared at the outset of the Plan. Allow for comments in relation to each presentation, but make sure to keep the time so that everyone can get comments. The comments – by the participants and you – should be oriented towards helping the author improve on their Plan. Note active participation for assessment purposes. Deliver your feedback and assessment either orally during the seminar or in writing after the seminar.

Assessment

The assessment of the course as a whole should be based on the learning outcomes and the assignment criteria.

Seminars: Keep track of participation and activity in the seminars

Plan: Assess that the Plan based on the assignment criteria. Language and every piece of information or every argument and every bibliographic reference (if there are any) do not need to be entirely perfect in order for the Plan to be acceptable, but as a whole it should demonstrate that the participant has understood the task, grasped the main contents of the three modules, is able to choose and define a realistic empirical material, suggest strategies, methods and tools for digitizing it, and state the intended purpose(s), outcomes, and user group(s) of the digitization project. Further, the Plan should not contain too obvious misunderstandings of literature, technologies, empirical materials or impossible proposals. If this is the case, so that the Plan does not reach the level of a Pass, the author should be allowed to revise it. Try to provide at least some comments to each author, pointing out interesting approaches or solutions as well as any problematic suggestions.

Both assignments - the minor on OCR and the major as Digitization Plan - are graded with a Pass or a Fail, and the participant needs to reach a Pass on both in order to receive a Pass on the course as a whole.

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



How to update the material

The course is based partly on content that should be relevant for the foreseeable future, and partly on content that is relevant only as historical or contemporary examples of applied technologies and solutions to particular problems. The needs, problems, solutions and technologies will for sure vary and be different for the participants of the course. If for instance a different software or format for text capture/OCR or text encoding is more relevant for the participants than those exemplified in the course, it would be a wise move to try to accommodate for this by slight modifications to the course content and its examples.

Some web sites that may be useful in keeping up-to-date and updating the course content are:

- Minerva EC <<http://www.minervaeurope.org>>
- Europeana <<http://www.europeana.eu>>
- The Text Encoding Initiative Consortium <<http://www.tei-c.org>>
- The Alliance of Digital Humanities Organizations <<http://adho.org>>
- *TEI by Example* <<http://www.teibyexample.org>>
- w3schools <<http://www.w3schools.com>>

And some useful journals:

- Digital Humanities Quarterly < <http://www.digitalhumanities.org/dhq/>>
- D-Lib Magazine <<http://www.dlib.org>>
- Liber Quarterly <<http://liber.library.uu.nl/index.php/lq>>
- Literary and Linguistic Computing <<http://llc.oxfordjournals.org>>
(subscription access).

Preparations for trainers

Prepare before the start of the course by glancing through the required reading, the recorded lectures, and some of the suggested additional readings, such as digitization guidelines.

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



Take a peek at the exercises and at both the assignments to learn what is expected from the participants (and to some extent, from you)

Familiarize yourself in due time with the learning platform so you can navigate it with ease and know what it contains, specifically how to upload and download material, and how to post in discussions and edit the contents. See if you are content with the prepared discussion threads for the course on the platform, or if you can foresee the need for additional threads and topics.

As soon as you have the list of participants before you, map where they are coming from, see if you can find out their affiliation - are they e.g. housed in libraries, and do you know if these are engaged in finished or ongoing digitization projects? The participants should be encouraged to tell a little about themselves and their particular interests and needs in the discussion forum of the learning platform at the start of the course. Make them describe whether or not they have experience from or are presently engaged in digitization projects or similar. What empirical material are the attendants working with and interested in? What kinds of questions are relevant for them?

If possible, familiarize yourself with cultural heritage digitization infrastructure, initiatives, and history of the country/countries in which the course participants are active as professionals.

Instructions for course designers

The introduction, the text about each module, the instructions for the two assignments, the digitized scans for the minor assignment on OCR, and a few other texts are made into separate pages on the learning platform. This would result in the following pages:

- Introduction [including Learning Outcomes and Course overview]
- Module 1: Digitization, purposes and strategies
- Module 2: Technology: text capture and encoding
- Minor assignment: OCR
- A page providing links to 8 scans to download for the minor assignment

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.



- Module 3: Use and assessment
- Major assignment: Digitization Plan
- Course literature and recommended literature [these are provided for and within each module, but all references could preferably also be collected on a separate page]
- Course assessment
- Course evaluation [assuming there is a standard for evaluating the courses – an evaluation has not been provided in this package]

Also, the following should be included:

- A discussion forum for participants and trainer, including these discussion threads [if possible, the trainer should also be able to add new threads]:
 - General course issues
 - Module 1
 - Module 2
 - Module 3
 - The minor assignment, OCR
 - The major assignment, Digitization Plan
- A possibility for participants to share documents with each other
- A place for participants to upload their major assignments (and possibly exercises) or to provide links to them. Upload formats should be allowed to vary, such as e.g. DOC, ODT, RTF, XLS, PDF, PPT, HTML

If the course is offered with webinars, participants need access to a video conferencing system with the possibility of displaying documents (e.g. PowerPoint presentations).

It would be advantageous if the URL:s are made into links to the external sites, and opened in a new tab or window. NB! Even though all the external educational content (and most of the recommended additional material, too) is available online free of charge in some version, most of it is covered by copyright restrictions.

All participants can have access to all the content from the beginning.

This project has been funded with support from the European Commission.

This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

