D4.1 Recommendations for metadata and data formats for online availability and long-term preservation

Succeed

16/01/2014

Abstract

This deliverable was prepared as part of the WP4 of the Succeed project. The objective of WP4 is to support the EC in the scope of activities identified in the Digital Agenda for Europe, by recommending a set of guidelines, formats, standards and licenses for digitization activities, both in terms of data and tools. The aim is to facilitate the implementation of digitization activities in the European institutions, by making the necessary tools and resources more interoperable, easily accessible and usable. This report provides a set of recommendations on formats and standards for digitization related activities, especially in the context of text/printed materials with focus on online delivery of digital objects and long-term preservation.

Succeed is supported by the European Union under FP7-ICT and coordinated by Universidad de Alicante.

Document information

Deliverable number	D4.1	Start: M3	Due: M12	Actual: M12
Deliverable name	Recommendations for metadata and data formats for online availability and long-term preservation			
Internal/External	External			
Activity type	SUPP			
Participant	KB, INL, PSNC, USAL, BVC, BnF, BRITISH LIBRARY			
Estimated person months per participant for this deliverable	As stated in Description of Work			
Dissemination level[1]	PU			

Document history

Revisions

Version	Status	Author	Date	Changes
0.1	Draft	Tomasz Parkoła	2.12.2013	Initial content
0.2	Draft	Tomasz Parkoła	9.12.2013	Section 2, 3 ready. Subsections 4.1-4.2 ready.
0.3	Draft	Jean-Philippe Moreux	10.12.2013	Improvements.
0.4	Draft	Tomasz Parkoła	13.12.2013	Sections 1,4,6 ready. Document ready for internal review.
0.5	Draft	Tomasz Parkoła	18.12.2013	Remarks from WP4 participants and Sebastian Kirch (internal supervisor) incorporated.
1.0	Final	Tomasz Parkoła	23.12.2013	Additional remarks from WP4 participants incorporated. Small improvements.
1.1	Final	Tomasz Parkoła	16.01.2014	Remarks from Jean-Philippe added.

Approvals

Version	Date of approval	Name	Role in project	Signature
0.4	16.12.2013	Sebastian Kirch	Supervisor D4.1	

Distribution

This document was sent to:

Version	Date of sending	Name	Role in project

[1] PU Public; RP Restricted to other programme participants (including Commission Services); RE Restricted to a group specified by the consortium (including Commission Services); CO Confidential, only for members of the consortium (including the Commission Services)

0.1	2.12.2013	WP4 participants	Project partners
0.3	10.12.2013	WP4 participants	Project partners
0.4	13.12.2013	Sebastian Kirch	Supervisor D4.1
0.4	13.12.2013	WP4 participants	Project partners
0.5	18.12.2013	WP4 participants	Project partners
1.0	23.12.2013	WP4 participants	Project partners
1.0	23.12.2013	Isabel Martinez	Succeed Project Technical Manager
1.0	23.12.2013	Marcel Watelet (EC)	Project officer
1.1	16.01.2014	WP4 participants	Project partners

Table of Contents

1. INTRODUCTION

This report provides a set of recommendations on formats and standards related to digitization activities, especially in the context of text/printed materials. The recommendations were identified based on existing recommendations, ongoing and emerging activities as well as an analysis of a dedicated survey conducted among digitization practitioners from various institutions across the world, including museums, libraries and archives. The aim of this report is to provide a set of recommendations, which are aligned with best practices of the most active digitization practitioners, especially those coming from Europe. The report also points out recommendations related to emerging standards and approaches as well as best practices, which are not entirely identified in the current digitization related activities.

This report is composed of four main parts, including a summary of existing recommendations, a summary of related ongoing and emerging activities, an analysis of the conducted survey and a set of Succeed project recommendations. Section 2 (existing recommendations) is intended to provide an overview of formats, standards and practices already applied in the digitization-related field. Each item described in this section has been supplemented with a summary table, providing the most valuable information to be considered when elaborating Succeed project recommendations. Altogether 17 existing recommendations or practices have been analyzed. Section 3 (ongoing and emerging activities) provides an overview of ongoing and emerging activities, which are the most interesting in the context of digitization. Described items include recent projects, innovative solutions and good practices that can contribute to Succeed recommendations, by showing mature and new approaches implemented or being implemented. Topics covered by this section include semantic technologies, OCR and linguistic resources, activities of relevant European Research Infrastructures as well as tools packaging. Section 4 provides a set of Succeed project recommendations, including those related to long-term preservation, online delivery of information as well as advanced and supporting technologies that can be used to improve and innovate digitization. Finally, the last section of this document is a summary and provides a concise view on Succeed recommendations, including overview tables and conclusions.

It is important to note that authors of this report assume that readers have basic knowledge of digitization-related concepts, such as metadata types (descriptive, structural, administrative), file types (master files, production files, delivery files), OCR, etc.

This report has been prepared in the framework of Succeed project, which is supported by the European Union under FP7-ICT and coordinated by Universidad de Alicante.

2. EXISTING RECOMMENDATIONS

This section summarizes existing recommendations and practices, which are related to digitization. Investigated aspects include metadata and data formats for online delivery and long-term archival of printed/text materials. For each item described in this section a summary table is provided. Such a table indicates the most important formats and standards that are recommended by described item. The idea of this section is to provide an overview of existing recommendations and practices, which can be taken into considerations during elaboration of Succeed project recommendations. The last section of this chapter provides statistics of all items described in this section. None of the items described in this section is older than 2004.

2.1 IMPACT project recommendations

The IMPACT project provides recommendations[2] on creation and management of metadata and images, including considerations related to image file formats. Therefore in case of metadata, specific formats and standards have been identified as candidates for use, while in case of file formats each candidate has also most important characteristics described. Table 1 presents a summary of the IMPACT project recommendations. The recommendations were prepared in the framework of the IMPACT project (years 2008-2012).

Table 1 Summary of IMPACT project recommendations

File Formats	
Master files	TIFF, JPEG2000, PNG
Delivery files	JPEG, JPEG2000, PNG, GIF
Metadata formats	
Descriptive	MARC, MODS, Dublin Core, EAD, TEI header
Structural	METS, MPEG-21 DIDL, OAI-ORE, TEI (textual content)
Administrative	textMD, NISO Z39.87 (MIX), PREMIS, NLNZ Preservation Metadata, LMER
Other formats	
OCR output	ALTO
Linguistic resources	
Tools packaging	
Other	

2.2 JISC Digital Media Guidelines

The JISC Digital Media Guidelines[3] provide an overview of factors that should be considered before choosing a file format, and suggest suitable file formats for specific applications. They also provide a comprehensive look at the various metadata standards

[2] http://www.digitisation.eu/fileadmin/user_upload/240/docbook/media/795a62a4-913d-ef04-1db7-3ac7ca3b28c0.pdf
[3] http://www.jiscdigitalmedia.ac.uk/guide/basic-guidelines-for-image-capture-and-optimisation

choices available to the developer of multimedia collections, and the principles behind using them. This summary is based on the sections related to choosing a file format for digital still images and metadata standards and interoperability. Table 2 presents a summary of the guidelines.

Table 2 Summary of JISC Digital Media Guidelines

File Formats	
Master files	DNG, TIFF, PNG, possibly PSD
Delivery files	JPEG, PNG, GIF, JPEG2000
Metadata formats	
Descriptive	MODS, Dublin Core
Structural	CMS, MPEG-21 DIDL, MPEG-21 for video
Administrative	METS Rights, MPEG RDD, PREMIS
Other formats	
OCR output	
Linguistic resources	
Tools packaging	
Other	

PSD is marked as a possible master file format. This is understood as an alternative to choosing whether to archive before or after optimization. The idea is to use the 'layers' features of Photoshop and save the image as a PSD file. This proprietary file format allows both the original image (un-optimized) and any optimization to be stored within the same file. The PSD file is however a proprietary format and its use should therefore be approached with great care.

2.3 Recommendations of the Bibliothèque nationale de France

The guidelines have been prepared by the Digitization Service of the Bibliothèque nationale de France (BnF). The current revision of the document is from November 8th 2013. The purpose of this publication is to document all requirements for image capture, metadata identification, OCR and ebook production, for the materials scanned as part of the BnF digitization programs. These guidelines and requirements are intended for service providers, institutions (libraries, archive centers) and others actors of cultural data digitization. The guidelines are relevant for manuscripts, books, graphic illustrations, artwork, maps, plans, photographs, objects and artifacts. Table 3 presents a summary of the guidelines.

Table 3 Summary of National Library of France recommendations

File Formats	
Master files	TIFF, JPEG2000 (2014)
Delivery files	Textual content: HTML, PDF with text layer, TXT; Still images: JPEG, JPEG2000
Metadata formats	
Descriptive	Proprietary format, METS (2014)

Structural	Proprietary format, METS (2014)
Administrative	Proprietary format, METS (2014)
Other formats	
OCR output	ALTO, PDF with text layer, TXT, TEI (navigation tables)
Linguistic resources	
Tools packaging	
Other	ePub

These guidelines ("référentiels" in French) are organized as several separate documents related to various tasks and themes: image digitization, text conversion, metadata identification, file delivery, etc. Important aspects in these themes are described below.

Image digitization

The guidelines suggest using an Adobe RGB ICC 98 color profile. All documents should be digitized using 24-bit depth except newspapers (8-bit greyscale). Seven resolutions are specified for different use cases, but the more common are 400 dpi and 600 dpi. TIFF V6 (monopage, uncompressed) is used as the raster image format for master files. JPEG2000 will be introduced in 2014 as the standard format for master files. The guidelines are available for opaque documents and transparent documents.

Text conversion (OCR)

The OCR guidelines focus on different tasks, including rules for OCR processing of documentary heritage (segmentation/structuring, recognition quality of textual contents), rules for implementing the ALTO format, quality control applied by the BNF QC team on contents produced (automatic control, visual inspection). A flavor of the ALTO LoC format is used, called "ALTO-prod"[4].

Text conversion (navigation table)

The navigation tables are used in the digital library website. They help readers to access to the digital content. The table guidelines focus on different tasks, including rules for page conversion of tables of contents and index in legacy documents, rules for structuration and transcription of these tables, rules of disqualification and simplifying of these tables, quality control applied by the BnF on the tables produced. The format for representing these navigation tables is an in-house XML format called "tdmNum". It's a XML schema based on TEI P4.

Text conversion (ePub)

The ePub guidelines focus on different tasks, including rules for converting legacy documents into digital book (editorial choices, technical requirements, etc.), requirements for the correction of the textual content, catalog metadata to be embedded in the ePub metadata, mechanisms used to improve accessibility (e.g. ePub 3 logical structuration), technical characteristics expected, technical and visual inspections

[4] http://bibnum.bnf.fr/alto_prod/documentation/alto_prod.html

performed by the BNF QC team. The ePub version described in the guidelines is ePub 3.0. The ePub version used in production is ePub 2.01.

Metadata

The metadata guidelines focus on different tasks, including rules for identification and description of heritage documents to create digital copies, rules for creating the lookup table physical document/digital document, rules for entering production data (types and levels of operations, dates, actors involved, hardware and software used, results found), rules for entering comments and captions. The format for representing these descriptive metadata is an in-house XML format called "refNum" (a METS flavor). METS will be introduced in 2014 in substitution of refNum.

Files delivery

The guidelines describe the architecture of the digital document package: folder names, hierarchy, etc. The package is a .zip archive, a .tar archive, or a zipped .tar archive. The formats for all the delivery files types are described in the other guidelines.

Diffusion formats

The guidelines don't suggest any specific format: the files are produced in-house (PDF, TXT, JPEG and JPEG2000). The only exception concerns the ebooks production, described in the ePub guidelines.

2.4 New York State Archives – Imaging Production Guidelines

The document[5] lists the minimal standards for producing and inspecting digital images of records. Table 4 presents a summary of the guidelines. The guidelines were published in 2008.

Table 4 Summary of New York State Archives - Imaging Production Guidelines

File Formats	
Master files	TIFF
Delivery files	TIFF, JPEG, JPEG2000, PDF/A
Metadata formats	
Descriptive	
Structural	
Administrative	
Other formats	
OCR output	ASCII, Unicode, XML
Linguistic resources	
Tools packaging	
Other	

[5] http://www.archives.nysed.gov/a/records/mr_erecords_imgguides.pdf

Master images should be at minimum 200dpi, for greyscale 8-bit depth, for color 16-36-bit depth. The images should be uncompressed. Backup images can be compressed using latest ITU standard compression. If delivery files need to be different from the master files then other formats and compressions are allowed. In case of compression, one should maintain uncompressed record copies to ensure easy accessibility to the image over time. Delivery files can be compressed using non-proprietary, lossless compression algorithms. They should be scaled so most documents fit within the typical computer screen or window for the given application. For instance, a particular application may require documents be scaled to half their size or less to comfortably fit a screen.

All images cannot have proprietary headers. Image orientation should be upright (portrait or landscape orientation should be maintained).

2.5 The NARA Technical Guidelines for Digitizing Archival Materials for Electronic Access

The U.S. National Archives and Records Administration (NARA) Technical Guidelines for Digitizing Archival Materials for Electronic Access[6] define approaches for creating digital surrogates for facilitating access and reproduction; they are not considered appropriate for preservation reformatting to create surrogates that will replace original records. The Technical Guidelines presented here are based on the procedures used by the Digital Imaging Lab of NARA's Special Media Preservation Laboratory for digitizing archival records and the creation of production master image files, and are a revision of the 1998 "NARA Guidelines for Digitizing Archival Materials for Electronic Access", which describes the imaging approach used for NARA's pilot Electronic Access Project. The Technical Guidelines are intended to be informative, and not intended to be prescriptive. They provide a technical foundation for digitization activities, and a range of options for various technical aspects of digitization, primarily relating to image capture. Table 5 presents a summary of NARA guidelines. The guidelines were published in June 2004.

Table 5 Summary of NARA technical guidelines

File Formats	
Master files	TIFF
Delivery files	JPEG, JPEG2000, GIF, PDF
Metadata formats	
Descriptive	Dublin Core, MARC
Structural	METS
Administrative	
Other formats	
OCR output	
Linguistic resources	

[6] http://www.archives.gov/preservation/technical/guidelines.pdf

Tools packaging	
Other	

For master files or production files use TIFF version 6, with Intel (Windows) byte order. Uncompressed files are recommended, particularly if files are not actively managed (e.g. stored on CD-ROM or DVD-ROM). If files are actively managed in a digital repository, it is possible to consider using either LZW or ZIP lossless compression for the TIFF files. JPEG compression should not be used within the TIFF format. DPI should be depended on Quality Index[7] (QI) and it should be equal to 8. It means that 600dpi (1-bit color depth) should be used for documents with smallest character of 1.0mm and 400 dpi in case of 8-bit greyscale images (also with smallest character of 1.0mm). For color images 24-bit depth in RGB mode should be used and 400dpi with smallest character of 1.0mm.

Access files should have sRGB or Adobe 1998 color profile and gamma 2.2 for greyscale.

2.6 NISO Framework of Guidance for Building Good Digital Collections

Developed by the National Information Standards Organization (NISO) in December 2007, this framework[8] aims to provide an overview of some of the major components and activities involved in creating good digital collections, to identify existing resources and to encourage community participation in the ongoing development of best practices. It also includes an extensive overview of existing guidelines and recommendations. Table 6 summarizes recommended practices indicated in the guidelines. The summary is focused on textual documents, therefore for example a/v content is not present there.

Table 6 Summary of NISO Framework of Guideance for Building Digital Collections

File Formats	
Master files	Textual content (structured format): TEI, TEI-lite, XML, PDF/A, PDF, ODF, SGML, Still images: TIFF, JPEG2000
Delivery files	Textual content: HTML, PDF; Still images: JPEG, PDF, GIF, JPEG2000, DjVu, MrSID
Metadata formats	
Descriptive	Dublin Core, MARC, MODS, ObjectID
Structural	METS
Administrative	copyrightMD, MIX, PREMIS
Other formats	
OCR output	TEI, TEI-lite, PDF
Linguistic resources	
Tools packaging	
Other	

[7] http://www.clir.org/pubs/abstract//reports/pub53
[8] http://www.niso.org/publications/rp/framework3.pdf

The framework established principles for digital collections, objects, metadata and initiatives and derives recommendations based on existing guidelines. In General, "digitals objects should exist in a format that supports its intended current and future use. It is therefore exchangeable across platforms, broadly accessible and formatted according to a recognized standard or best practice".

Quality recommendations for the digitization process vary: Most common is a bit depth of 8 per channel or 24 bits per pixel and a spatial resolution of 300 to 600 dpi. Exceptions like very old manuscripts may require a resolution of up to 2400 dpi. The master file format is usually uncompressed TIFF, which is very well established, with a growing interest in employing JPEG2000 images as masters or archival formats. For end-users, formats such as GIF or JPEG can be used or even created on-the-fly.

Derived texts for search and retrieval should be provided as marked-up texts within an established XML schema or DTD such as SGML, TEI or TEI-lite. PDF and PDF/A are other options. HTML is acceptable for publication and dissemination. Generally, textual content should be represented in a way that it can be accessed by search engines. The encoding should be either US-ASCII or UTF-8.

"Good metadata conforms to community standards in a way that is appropriate to the materials in the collection, users of the collection, and current and potential future uses of the collection." Therefore, one should make use of well-established metadata schemes, controlled vocabularies and thesauri. The recommendations include a variety of metadata schemes for different objects and domains. For cultural heritage institutions, the most important schemes are Dublin Core, MARC21 and MODS for descriptive metadata, METS for structural metadata and copyrightMD and MIX for administrative metadata.

2.7 California Digital Library Guidelines for Digital Objects and File Format Recommendations

These guidelines[9] were developed by the California Digital Library (CDL) in August 2011. They provide specifications for digital objects prepared by the institutions for submission to CDL. Table 7 summarizes recommended practices indicated in the guidelines. The summary is focused on textual documents, therefore for example a/v content is not present there.

Table 7 Summary of California Digital Library guidelines

File Formats	
Master files	Textual content (structured format): PDF/A, HTML, XML, TXT (UTF-8 or ASCII), TEI, ALTO; Still images: TIFF, JPEG2000
Delivery files	Still images: JPEG, GIF, PNG

[9] http://www.cdlib.org/gateways/docs/cdl_dffr.pdf
http://www.cdlib.org/services/dsc/contribute/docs/GDO.pdf

Metadata formats	
Descriptive	Dublin Core, MARC, MODS
Structural	METS
Administrative	METS Rights extension schema, MIX, PREMIS
Other formats	
OCR output	TEI, ALTO
Linguistic resources	
Tools packaging	
Other	

The guidelines seek to ensure a basic level of uniformity in the interoperability, management, structure and encoding of digital content managed by the CDL. Therefore, they rely on formats that are well supported and are more likely to be accessible in the future. The guidelines define basic criteria for these formats:

- Non proprietary
- Open, documented standards
- In common usage by the research community
- Use standard character encoding (ASCII, UTF-8)
- Unencrypted
- Uncompressed

Digital objects are categorized in three classes: metadata, content files and a link or binding mechanism to associate the two. For each of these classes the guidelines suggest particular formats and procedures.

CDL offers different services and the requirements/recommendations depend on the service used. In general, metadata is managed using the METS format utilizing METS profiles. Each METS file submitted must conform to valid METS profiles. Generally it is advised to provide the most granular and richest metadata possible using schemas such as MODS instead of just simple or qualified Dublin Core (A guideline for descriptive metadata elements is provided). UTF-8 or UTF-16 should be used for character encoding. Technical Metadata is derived from the digital objects using JHOVE. Any additional technical metadata is optional but should be encoded using valid XML extension schemas such as the NISO Metadata for Images in XML Schema (MIX).

Recommendations for content files are further elaborated in the CDL DFFR guidelines. For graphical production master files it is advised to use uncompressed TIFF. Color and grayscale files should have ICC color profiles embedded in the file header. Display or thumbnail images are provided using the JPEG, GIF or PNG format. Images should be 800 – 3000 pixels across the long dimension and have a medium or high compression. For text files the guidelines rely on the PDF/A and the TEI standard. It is recommended to include embedded text transcriptions in PDF files when possible and it is advised to submit one PDF or TEI file per digital object. Alternative options for text file formats are HTML, XML and TXT files. For full-text transcriptions, ALTO can be used.

2.8 DFG-Praxisregeln "Digitalisierung" [DFG guidelines on digitization]

The DFG guidelines[10] on digitization aim to ease the application for digitization project funding through the DFG by providing best practices and common standards. Table 8 summarizes recommended practices indicated in the guidelines. The summary is focused on textual documents, therefore for example a/v content is not present there. The guidelines were published in February 2013.

Table 8 Summary of DFG guidelines

File Formats	
Master files	TIFF, JPEG2000
Delivery files	JPEG, PNG
Metadata formats	
Descriptive	MOTS, TEI, LIDO, EAD, Dublin Core Collections Application Profile
Structural	
Administrative	PREMIS
Other formats	
OCR output	ALTO, PDF
Linguistic resources	
Tools packaging	
Other	

The DFG guidelines on digitization are very extensive and cover many aspects of the digitization process including file formats, technical equipment and organizational questions. However the recommendations on file formats for images and metadata are rather straightforward to achieve a high level of uniformity in digitization projects.

Master files should be scanned with a spatial resolution of at least 300 dpi and a bit depth of 8 bit per channel per pixel (24bit in all). It is recommended to use the uncompressed TIFF file format and not to use extended TIFF variants such as Baseline-TIFF. Lossless JPEG2000 is also a feasible option for master images. However, JPEG, PNG or proprietary formats like vendor-dependent RAW formats should not be used. For publication and dissemination the guidelines advise to use JPEG or PNG as the file format of choice as they are the most widely adapted standards.

Metadata should be provided in a software-independent and standardized format such as XML. It is very important that the creation of metadata is deeply integrated into the production workflow to ensure that even if a project is aborted metadata is available for those objects that have been processed so far. Usually METS is the container format of choice with embedded metadata in formats such as MODS or TEI. By using formats like LIDO or EAD it is possible to reference additional external objects such as audiovisual content. In summary, METS/MODS should be used for printed texts, METS/TEI for manuscripts, EAD or SAFTXML for archival material and LIDO for pictures or three-

[10] http://www.dfg.de/formulare/12_151/12_151_en.pdf

dimensional objects. To ensure the sustainability of descriptive metadata, the guidelines advise to use standards and reference models like CIDOC CRM or Dublin Core. References to the content files always need to be integrated into the metadata file. Additionally, metadata must be made available via OAI.

OCR is an important part of the digitization workflow and the accuracy of OCR results should always be verified. ALTO is the format of choice for storing OCR output (UTF-8 encoding). Additionally, a PDF can be provided.

2.9 The Getty Research Institute online resources

The Getty Research Institute makes available an online publication[11] that introduces the technology of digital imaging and creating an image library. The publication provides an introduction to photographers, publishers and memory institutions on imaging, but does not contain real recommendations. Table 9 provides a summary of information related to this online publication. The resources were published in 2008.

Table 9 Summary of the introduction to imaging and metadata (The Getty Research Institute)

File Formats	
Master files	TIFF, PNG, JPEG2000
Delivery files	JPEG
Metadata formats	
Descriptive	MARC, MODS, CDWA, CIDOC CRM, SPECTRUM, Dublin Core
Structural	METS
Administrative	MIX
Other formats	
OCR output	
Linguistic resources	
Tools packaging	
Other	

2.10 Universal Photographic Digital Imaging Guidelines

The UPDIG Image Receiver Guidelines[12] recommend standards to improve the "hand-off" of digital image files from photographers to end users of all types. This diverse community includes stock image distributors, magazine and book publishers, publication designers, web designers, art directors, museums, printers, fine-art publishers and more. UPDIG's Digital Imaging Submissions Guidelines working group reviewed the submission practices of various end-user communities, identified best practices and standards that meet their needs, and produced this document to help users develop their own Submission Guidelines. Members of the DISG working group represent both digital

[11] http://www.getty.edu/research/publications/electronic_publications/intrometadata/index.html
http://www.getty.edu/research/publications/electronic_publications/introimages/index.html
[12] http://www.updig.org/index.html

image suppliers and user communities, including magazine publishers, stock image distributors, graphic and web designers, museums, and others.

There are a number of principles that form the foundation of these guidelines:

- Digital images should look the same as they transfer between devices, platforms and vendors.
- Digital images should be prepared in the correct resolution, size and sharpness for the device(s) on which they will be viewed or printed.
- Digital images should have embedded metadata that conform to the IPTC and PLUS standards, making them searchable while providing relevant rights and usage information – including creator's name, contact information and a description of licensed uses.
- Digital images should be protected from accidental erasure or corruption and stored carefully to ensure their availability to future generations.

Summary of the recommendations is presented in the Table 10. The guidelines were published in December 2008.

Table 10 Summary of Universal Photographic Digital Imaging Guidelines

File Formats	
Master files	RAW, DNG,
Delivery files	JPEG, TIFF, JPEG2000
Metadata formats	
Descriptive	EXIF, IPTC core schema, PLUS, XMP
Structural	
Administrative	
Other formats	
OCR output	
Linguistic resources	
Tools packaging	
Other	

2.11 Federal Agencies Digitization Initiative Guidelines

Federal Agencies Digitization Initiative Guidelines cover various aspects of digitization. One of them is the online publication related to technical guidelines for digitizing cultural heritage materials focused on creation of raster image master files. The current revision of the document is from August 24th 2010. The guidelines are based on the National Archives and Records Administration's Technical Guidelines for Digitizing Archival Records for Electronic Access: Creation of Production Master Files – Raster Images (June 2004). The guidelines are relevant for manuscripts, books, graphic illustrations, artwork, maps, plans, photographs, aerial photographs, objects and artifacts. Summary of the guidelines is presented in the Table 11.

Table 11 Summary of the Federal Agencies Digitization Initiative Guidelines

File Formats	
Master files	TIFF, JPEG2000, PNG
Delivery files	JPEG
Metadata formats	
Descriptive	Dublin Core
Structural	METS
Administrative	MIX
Other formats	
OCR output	
Linguistic resources	
Tools packaging	
Other	

The guidelines suggest to use Adobe RGB 1998 color profile or sRGB color profile for digital still images, as currently the safest way to store information of color mode due to the popularity (comparing to LAB color profile) and rich color gamut (comparing to CMYK color profile).

Textual documents should be digitized using 1-bit, 8-bit or 24-bit depth. It is indicated that 8-bit greyscale can be best performing for older textual documents with low contrast (e.g. where background and text does not differ much or halftone images are embedded into text). For the images where the color is important 24-bit is obviously necessary. Specific parameters of images, including Quality Index[13] (QI) parameter are outlined in the guidelines. It especially concerns dpi and color depth.

In the context of master files TIFF is presented as "de facto" standard for raster image format. A number of technical characteristics are presented in regard to TIFF, including XMP support, various color spaces and profiles support, long track record (format has over 10 years) and failure to directly use as delivery format (unsupported by web browsers). JPEG 2000 is presented as increasingly considered format for master files, but is not yet widely adopted. JPEG2000 characteristics include a complex model for encoding data (in comparison to TIFF), multiple resolution and color profiles support and extensive XMP support. PNG is indicated as a format which is applicable for production master files, although it is not commonly in use. The main characteristics include simplicity, lossless compression, later web browsers native support. The final recommendations state that TIFF is the best option for master files. It should be TIFF version 6 with Intel byte order and preferably uncompressed. If necessary ZIP compression should be used (not JPEG).

[13] http://www.clir.org/pubs/abstract//reports/pub53

In case of delivery files JPEG/JFIF is recommended for items other that text or line drawings. GIF on the other hand is recommended as an access format for textual documents, but limitation to 8-bit color depth needs to be taken into consideration.

PDF, PDF/A, ASCII and XML are presented on the formats list considered in the guidelines, but no clear guidelines are present. PDF is presented as a highly structured page description language with XML support, limited color spaces support and various compression possibilities of different parts of the file. PDF is not recommended as production master file format. ASCII is considered to hold the text of image files, but at the same time it means loss of structure and formatting. XML is again considered as text holder with support for building text hierarchy, simple to use in the context of retrieval and interoperable format. No recommendations are specified in the context of ASCII and XML.

In the context of metadata it is underlined that the guidelines are focused on discussing different metadata formats rather that recommending specific ones. This is due to the fact that different projects need different metadata elements to describe digital content in the required details.

In case of descriptive metadata Dublin Core is suggested as format for representing minimal descriptive metadata. It should be collected either directly when no metadata is available for an item, or via mapping when there is already a local metadata schema present.

Administrative metadata are divided into several items based on their function, which are: Rights, Technical, Behavior, Preservation and Tracking. In the context of metadata related to all functions several options are mentioned (e.g. extensions to METS), but no specific format is recommended. For rights metadata no clear guidance is provided, it is rather in form of options to be considered. In case of technical metadata FADGI recommendations refer to ANSI/NISO Z39.87 Data Dictionary - Technical Metadata for Digital Still Images, for which an XML schema has been developed at the Library of Congress – MIX. It is also indicated that registries like Global Digital Format Registry could help in a way that technical metadata can reference to it, not needing to embed all the information available in the registry. For preservation metadata several options are discussed, e.g. PREMIS.

For structural metadata METS is discussed and recognized as format useful in the library context, rather than in archival as METS does not apply well when it comes to hierarchical collections.

2.12 Technical Guidelines for Digital Cultural Content Creation Programmes (MINERVA project)

The guidelines have been created in the context of various European and national initiatives, by MINERVA EC project working group and NRG activities. Table 12 summarizes recommended practices indicated in the guidelines. The summary is focused

on textual documents, therefore for example a/v content is not present there. The guidelines were published in September 2008.

Table 12 Summary of the Technical Guidelines for Digital Cultural Content Creation Programmes

File Formats	
Master files	Textual content: structured format such as XML, TEI), PDF/A; Still images: TIFF, PNG, GIF, JPEG (SPIFF); Graphics (raster): PNG, GIF; Graphics (vector): SVG, Macromedia SWF
Delivery files	Textual content: XHTML 1.0, HTML 4 or newer, PDF, ODF, RTF; Still images: JPEG (SPIFF); Graphics (raster): GIF, PNG; Graphics (vector): SVG
Metadata formats	
Descriptive	Dublin Core, Dublin Core Collections Application Profile, NISO metasearch collection description specification, MICHEL Data Model
Structural	METS, IMS CPS
Administrative	NISO Z39.87-2002, PREMIS, Creative Commons
Other formats	
OCR output	
Linguistic resources	
Tools packaging	
Other	

In general the guidelines suggest using open standard formats for creating digital resources in order to maximize the access and increase the interoperability. Proprietary formats are acceptable in cases where there is no other option, but then migration strategy for those resources should be considered.

In case of all textual data formats (both master and delivery) it is recommended to explicitly state text encoding.

In case of master files for still images it is stated that TIFF is most applicable for photographic images. Two parameters have been considered in relation to still images: spatial resolution and color resolution. In case of photographic prints it is recommended to use 600 dpi and 24-bit depth for color or 8-bit depth for grayscale. In case of 35mm slides it is recommended to consider 2400 dpi resolution.

Graphic non-vector images are those produced in digital form, which represent logos, icons and line drawings. It is recommended to use 72dpi for recording master files of those items. No specification on dpi is given for delivery files.

For vector images it is suggested to use SVG – an open standard for representing vector images. It relates both for master files and delivery files. In some cases it is possible to use proprietary Macromedia SWF for storing master files.

It is recommended that delivery files for still images are provided according to EMII-DCF, which means that images for full screen should be provided at 150dpi and with 24-bit color or 8-bit grayscale, using a maximum of 600 pixels for the longest dimension (this resolution is lower that required for high resolution prints). Their thumbnails should be provided at a resolution of 72 dpi, using a bit depth of 24-bit color or 8-bit greyscale, and using a maximum of 100-200 pixels for the longest dimension.

In case of administrative metadata (includes preservation, technical, provenance, rights metadata) the guide does not provide clear recommendations (except rights metadata) – it rather lists possible options and suggest to record all sufficient metadata needed to manage institutional digital resources. These options are listed in the Table 12 above (in administrative metadata formats section). For rights metadata it is recommended to clearly state what the possibilities in terms of reuse of metadata and digital content are. Creative Commons licenses are recommended.

In case of structural metadata IMS CPS is recommended to use when working with learning resources, while METS is suggested to be used for other digital objects.
Descriptive metadata should be specified in a format, which is based on Dublic Core simple/unqualified. For collection level description several options should be considered (Dublin Core Collections Application Profile, NISO metasearch collection description specification or MICHAEL Data Model). It is also suggested to take advantage of the metadata terminology referenced in the MICHAEL Data Model (collection level description). For learning objects it is recommended to consider IEEE LOM (learning object metadata).

The aspects of semantic web and ontologies are also discussed, but without clear recommendations of the format. It is rather listing of available formats, which can be used to leverage semantic technologies.

2.13 The National Digital Newspaper Program - Technical Guidelines for Applicants

The National Digital Newspaper Program (NDNP) is an effort to build a database of U.S. newspapers by mean of an award programme. NDNP is a partnership between National Endowment for the Humanities (NEH) and Library of Congress (LoC), where NEH provides funds for digitization and LoC assures permanent maintenance. The partnership agreed on technical requirements for applicants. The approach is to assure long-term and short-term goals, including accessibility via WWW, good image quality for OCR and appropriate approach in the context of long-term preservation. In general NDNP follow Federal Agencies Digitization Guidelines Initiative (FADGI). It means that recommendations are based on the options discussed in FAGDI documents related to still images. Summary is presented in the Table 13. The guidelines were published in 2013.

Table 13 Summary of the National Digital Newspaper Program Technical Guidelines for Applicants

File Formats	
Master files	TIFF, JPEG2000

Delivery files	PDF with text layer
Metadata formats	
Descriptive	MARC, MODS, Dublin Core
Structural	METS
Administrative	MIX, PREMIS
Other formats	
OCR output	ALTO
Linguistic resources	
Tools packaging	
Other	

Master files should be stored in TIFF format, version 6.0, uncompressed. The images should be digitized in greyscale using 8-bit color depth. It is recommended have the maximum possible resolution, which means 300-400 dpi depending on the physical dimensions of the original. TIFF tags should include selected descriptive and technical metadata. All the master files should be also provided in JPEG2000 format, so that it can be accessed via dedicated user interface over the web based on JPEG2000 wavelet compression (zooming, segments). Each JPEG2000 needs to have XMP metadata, 6 decomposition levels and 25 quality levels with compression rate 8:1. JPEG2000 files are supposed to be derivatives from the TIFF files.

Delivery files should be provided in PDF format with hidden text, making it therefore a searchable document. PDF should include XMP metadata and it is recommended to conform to the PDF/A specification. Each page of such a PDF should be in JPEG format with 150dpi resolution (also greyscale).

Descriptive metadata are required to be provided in MARC21 communication formant using UTF-8 encoding. It should be provided prior to the upload of digital asset. In the digital asset itself the metadata should be provided in METS format, which contains metadata in the following formats: Dublin Core, MIX, MODS, PREMIS. Additionally there are specific metadata elements for NDPD programme, such as identifiers in scope of the Library of Congress as not all items have ISSN assigned.

OCR output should be recorded using ALTO version 2.0 or newer.

2.14 Images for web delivery – standards, image capture standards, metadata for images created by the National Library of Australia

The following image capture standards are used by the National Library in digitization of its collection material. A range of derivatives are produced for Web delivery of the National Library's digitized collection material. The National Library of Australia is progressively making information about its digitized collection materials available using the Open Archives Initiative (OAI) protocol for metadata harvesting. This service provides access to metadata describing the Library's digital collections, which is held in

Succeed is supported by the European Union under FP7-ICT and coordinated by Universidad de Alicante.

the Digital Collections Manager (DCM) database. The documentation was created in June 2004. Summary is presented in the Table 14.

Table 14 Summary of National Library of Australia practices

File Formats	
Master files	TIFF
Delivery files	JPEG, PDF, MrSID
Metadata formats	
Descriptive	
Structural	
Administrative	
Other formats	
OCR output	
Linguistic resources	
Tools packaging	
Other	

Master Files should have tonal resolution of 24 bits per pixel and 300 ppi spatial resolution for larger than A4, 600 ppi between A5 and A4 format, 1200 ppi between A7 and A6, and 2000 ppi under A7 format.

Thumbnail copies of derivative files are compressed with 72 ppi and with dimension of 150 pixels. View copy (JPEG) is created with Image Alchemy software with 72 ppi and longest dimension of 600 pixels for pictures and 760 pixels for manuscripts, maps and music. View copy (multi-page PDF) is created for print publications scanned for Copies Direct orders. These images are compressed with 72 dpi and longest dimension 1000 pixels. Examination copy (JPEG) for printed music and cartographic materials from TIFF master using Image Alchemy software with 72 ppi of resolution and with longest dimension 1000 pixels. Print copy (PDF) for printed music from JPEG examination copies using Image Alchemy software. These files are compressed with 72 dpi resolution and with longest dimension 1000 pixels. Interactive copies (MrSID) are created primarily for cartographic material from TIFF master using MrSID software. These files are compressed with 300 ppi resolution and with longest dimension as per the TIFF master (varies according to the original physical item).

2.15 University of Virginia Library – community digitization guidelines

The document offers guidance and minimum recommendation in line with UV Library's current practice. The guidelines are divided into two main topics: Digitization requirements and Metadata. The document was created in March, 6th 2006, but it claims that it is a continually evolving document. The summary is shown in the Table 15.

Table 15 Summary of University of Virginia Library community digitization guidelines

File Formats	
Master files	TIFF, JPEG2000

Delivery files	PDF with text layer
Metadata formats	
Descriptive	MARC, MODS, Dublin Core
Structural	METS
Administrative	MIX, PREMIS
Other formats	
OCR output	ALTO
Linguistic resources	
Tools packaging	
Other	

Master files should be stored in TIFF format, uncompressed with a resolution between 300 and 600 ppi and 8-bit (grayscale) 24-bit (color) depth, depending if they are text pages of a book or illustrations, slides or oversize items.

For access copies to the master files, the documents should be digitized in TIFF uncompressed, but for the text pages of a book that should have a CCITT Group 4 Fax compression, and with a resolution between 300 and 400 ppi, depending on the type of original. Delivery files should be provided in JPEG format with a resolution between 120 px and 3000 px on the longest side, depending if its purpose is thumbnail, screen-sized or maximum, automatically compressed (select High or level 10).

Electronic texts should be captured in XML, XHTML, ASCII text or PDF, depending on the purpose. The standard used in XML is TEI P4. The same formats will be used in deliverables.

Other formats, as summarized in the table above, are regarding video, audio, numeric and spatial data.

The document doesn't give any recommendation on metadata format, just the content they must include. This content is divided into required fields, recommended fields and optional fields.

2.16 Image Specifications and Functional Requirements for Citation Capture (PubMed Central Back Issue Scanning Project)

This section summarizes the specification of digitization parameters of the PubMed Central Back Issue Scanning Project by the National Library of Medicine. Table 16 summarizes recommended formats and deliverables indicated in the guidelines. The summary is focused on digitization parameters for textual documents. The document was created in May 2007.

Table 16 Summary of technical digitisation parameters for PubMed Central Back Issue Scanning Project

File Formats	
Master files	TIFF
Delivery files	TIFF, PDF
Metadata formats	
Descriptive	XML
Structural	
Administrative	
Other formats	
OCR output	ASCII
Linguistic resources	
Tools packaging	
Other	

The guidelines provide detailed instructions on an article based digitization workflow for the National Library of Medicine.

Digitized material is to be delivered organized at article (not page) level. The document therefore includes an extensive set of instructions on folder/file naming scheme and treatment of pages that have content belonging to more than one article.

In terms of image digitization, 2 different image types are identified: plain text pages which are to be delivered as 600 dpi bitonal TIFF as whole page scans and color or greyscale illustrations which are to be delivered as 300 dpi, 24 bit color or 8 bit greyscale TIFF, cropped to the size of the illustration.

Also to be delivered are article level PDF files containing the bitonal scans of the article pages as well as the OCR result as hidden text (for searching purposes). Finally, unedited OCR results using Prime OCR are to be delivered as plain text ASCII files for each article (not page).

In terms of metadata, a collection of files is also to be generated and delivered according to the digitization workflow described. An XML tagged article level citation is to be created (DTD for this file is supplied within the specification document).

Index files per media disk delivered and a file mapping inventory linking volume and issue numbers with the paths of generated files are also to de generated. The format for both these files is also described within the specification document.

2.17 Picture Queensland Image Digitisation Manual 2007

This document summarizes the specification of digitization parameters of the Image Digitization Manual by the State Library of Queensland. Table 17 summarizes recommended formats and deliverable indicated in the guidelines. The summary is

focused on digitization parameters. The document is providing guidelines for photograph digitization. The document was published in 2007.

Table 17 Summary of Picture Queensland Image Digitisation Manual 2007

File Formats	
Master files	TIFF
Delivery files	TIFF, JPEG
Metadata formats	
Descriptive	Dublin Core
Structural	
Administrative	
Other formats	
OCR output	ASCII
Linguistic resources	
Tools packaging	
Other	

The guidelines provide instruction and a step-by-step guide for photograph digitization for the State Library of Queensland. There is detailed explanation of planning a digitization process, although most of the guidelines provided are relevant only to photographic material.

As far as the digitization is concerned, the guidelines distinguish between two different types of photographs: black-and-white and color. For black-and-white originals, an 8-bit greyscale TIFF image is captured and for color originals a 24-bit color TIFF image is recommended. In either case, a minimum size (in pixels) is advised - 6,000 pixels for black-and white photographs and 4,000 pixels for color. This is achieved by altering the scanning resolution so that the resulting image satisfied the minimum size is pixels. Tables are provided for easier selection of scanning resolution based on the size (in inches) of the original.

Following the scanning, a number of manual image editing steps are described. These consist of simple rotation and cropping operations, but also include adjusting color levels, smoothening, resizing and finally adding noise (a 1% uniform noise filter is applied in order to "smoothen out sharp spikes in the levels histogram").
At the end of this manual process images are saved as TIFF and if required (it is not specified as a compulsory step), JPEG for access copies.

Finally, indexing metadata are entered per scanned image. The metadata is stored using Dublin core elements schema and there is a list of 15 fields that can be used to index each image (covering descriptive and administrative information.

2.18 Summary of existing recommendations

This section provides an overview of 17 items related to practices and recommendations implemented around the world. From the general perspective it is visible that current practices and recommendations do not cover topics related to the whole digitization workflow. For example only 11 items out of 17 have indicated OCR output formats and linguistic resources. Also tools packaging have not been mentioned at all.

The charts in this section present the percentage of recommendations/practices that indicate particular format as an option for use. Labels on the data columns indicate the number of recommendations mentioning particular format.

Figure 1 presents statistics for the master files formats. Over 94% of items described in this section suggest usage of TIFF format for master files. It goes along with the common understanding (in the digitization community) of the TIFF format as "de facto" standard. The second most common format indicated in the analyzed recommendations is JPEG2000. Although the format is quite complicated and still does not have wide adaptations, several institutions across the globe use JPEG2000 as an archival master file. PNG is indicated by less than 30% and usually it is understood as a format, which is not commonly used, therefore not the best option to be used. Because most of the practices or recommendations focus on still images, DNG format and others mentioned in the described items are not visible in the summary. Nevertheless it is important to mention that in case of photography DNG format may be considered, while in case of textual content XML-based formats are mostly indicated (e.g. TEI).

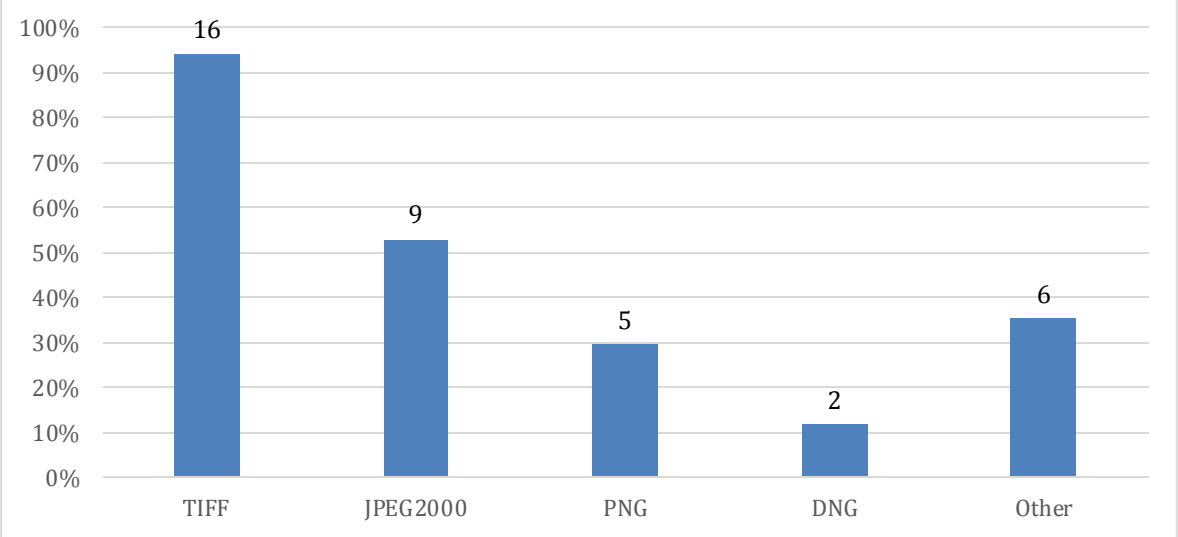

Figure 1 Summary for master file formats

The most important delivery format for all items in the analysis is JPEG (see Figure 2). It was indicated by more than 80% of items and is understood as a very good option for all types of delivery files, including presentation version and thumbnails. The other delivery formats mentioned by more than 25% of items include PDF, JPEG2000, GIF and

PNG. For PDF the application is mostly in the context of textual content, while in case of JPEG2000, GIF and PNG it is still images. TIFF format has been indicated as a delivery format in approx. 20% of cases, but because of its characteristics (e.g. browser support, compression scheme), it is rather not a good option to consider in this context. From the general perspective it seems to be most reasonable to use JPEG, PNG and GIF for still images and PDF for textual images. JPEG2000 seems to be a good alternative in the context of emerging formats. This is because it is able to provide both delivery file and master file in a single JPEG2000 file. In case of GIF it is important to remember its limitations (e.g. maximum of 256 colors per image) and usual use cases (animation and sharp-edged line art). PNG can serve both master files and delivery files, but as opposed to JPEG2000, it needs conversion to delivery format and therefore existence of two different files with different characteristics (e.g. resolution). PNG have issues with older web browsers and it is used mostly with lossless compression scheme, therefore usually gives larger files than JPEG. Because of several advantages of PNG over GIF it seems to be more reasonable to use PNG for still images rather than GIF. The advantages include number of colors or transparency options.

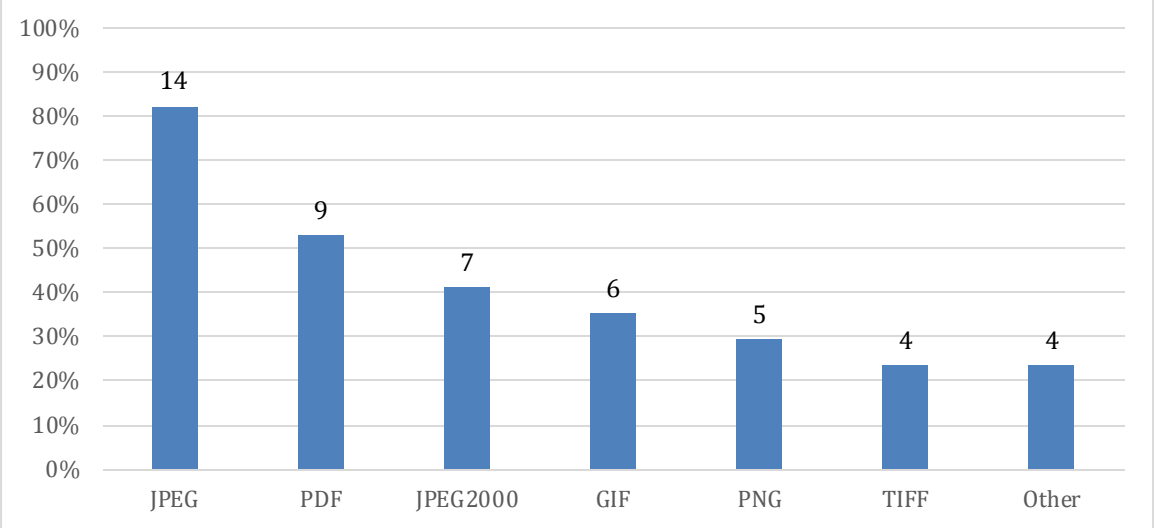

Figure 2 Summary for delivery file formats

For the purpose of analysis three types of metadata formats have been defined (according to the NISO standard[14]). These include descriptive metadata formats, structural metadata formats and administrative metadata formats. In case of descriptive metadata formats the recommendations and practices are mostly focused on XML formats, including Dublin Core and MODS. More than 40% of the items indicated MARC format. See Figure 3 for details.

[14] http://www.niso.org/publications/press/UnderstandingMetadata.pdf

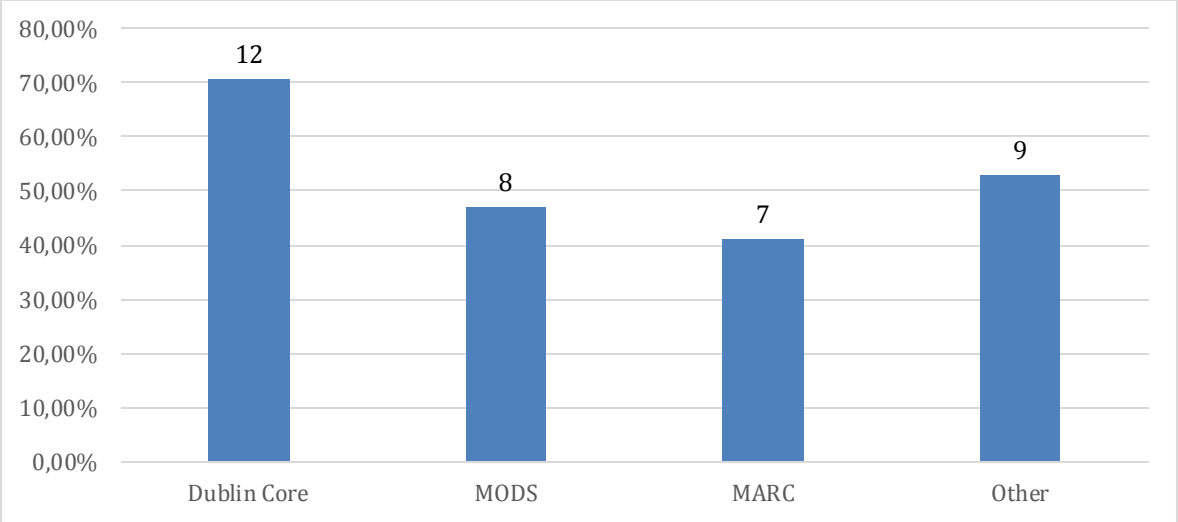

Figure 3 Summary for descriptive metadata formats

In case of structural metadata formats the most common selection is METS (pointed by almost 60% of items), and it is in fact the only one pointed by more than 10% of analyzed recommendations and practices (see Figure 4).

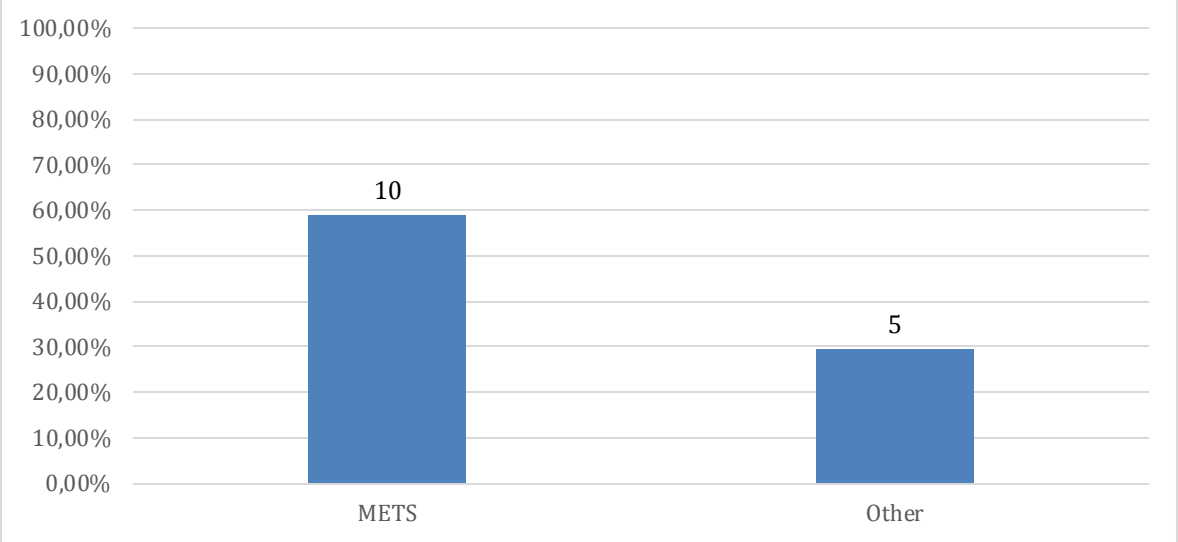

Figure 4 Summary for structural metadata formats

In case of administrative metadata the most popular standards and formats are related to technical and preservation metadata and include PREMIS as well as NISO Z39-87. MIX, which is an XML implementation of the NISO Z39-87 dictionary, is also a common selection in this context. Altogether MIX and NISO Z39-87 reach more than 45% of indications, while PREMIS reaches more than 40% (see Figure 5).

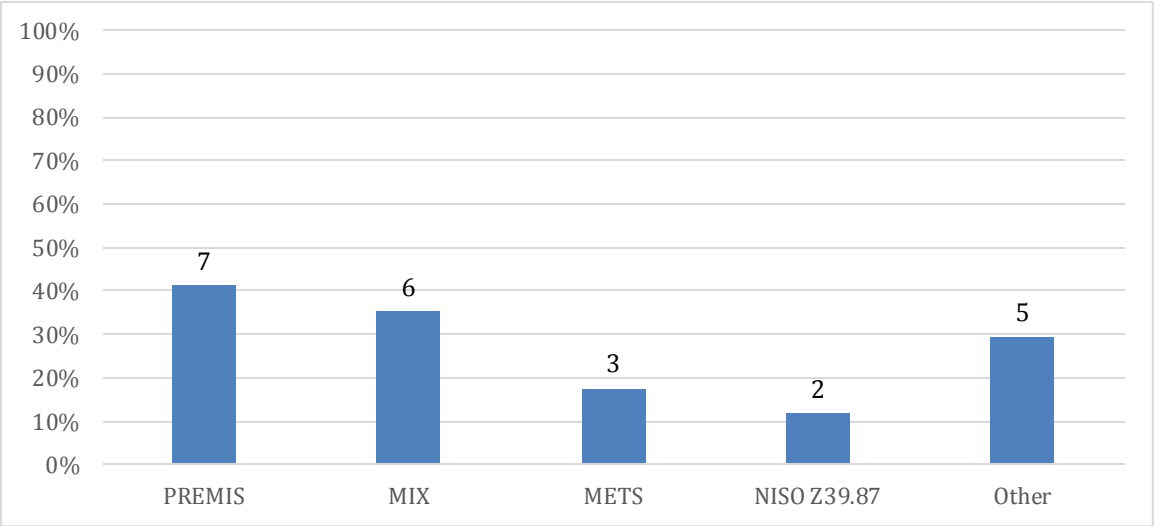

Figure 5 Summary for administrative metadata formats

For the OCR results representation most indications are related to ALTO format (almost 30%). The other common indications include ASCII format, Unicode and TEI. Because the ASCII and Unicode formats are quite popular (together more than 35% of indications), it is worthwhile of taking advantage of the UTF-8 encoding (of the Unicode character set), which is compatible with ASCII and Unicode at the same time. In the context of OCR results we need to remember that only 11 of the analyzed items indicated a format, the rest (6 items) do not tackle the issue of OCR results representation at all (see Figure 6).

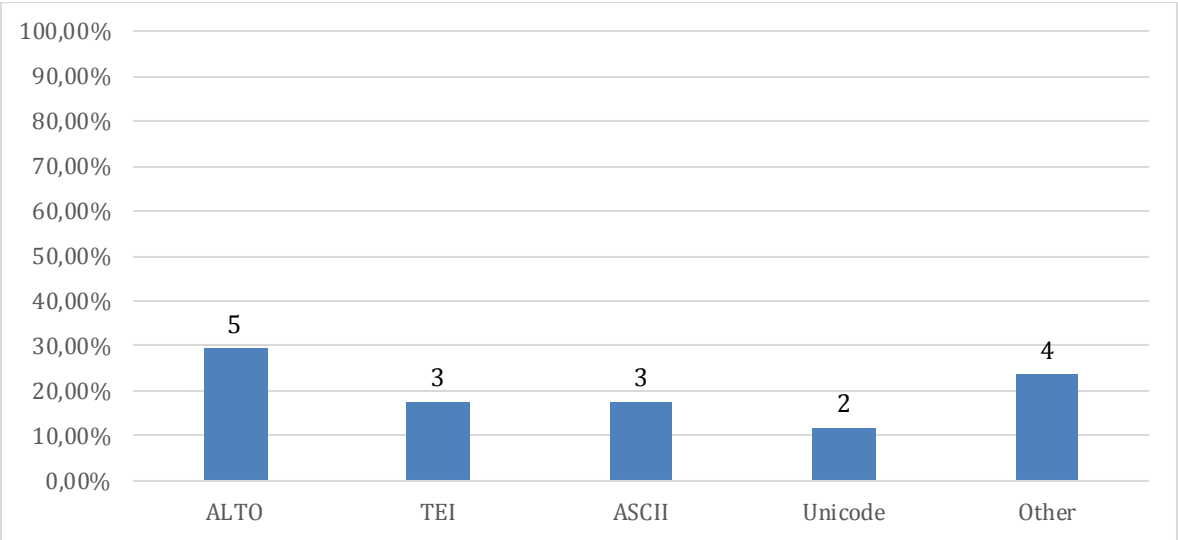

Figure 6 Summary for OCR formats

3. RELATED WORK – ONGOING AND EMERGING ACTIVITIES

This chapter presents ongoing and emerging technologies and formats that are used by various consortia, projects and initiatives in order to enhance digitization-related activities. The chapter covers semantic technologies, OCR and linguistic resources, relevant ERICs[15] as well as tools packaging issues.

3.1 Semantic technologies

It is already common understanding that semantic technologies play an important role in digitization-related activities. Various projects and initiatives (re)designs their data, so that it is possible to expose them with semantic technologies. The idea of Linked Open Data (LOD) has been especially investigated in this area. The following subsections provide an overview of the applications and usages of LOD in the cultural heritage context and digitization.

Digitised Manuscripts To Europeana

Introduction

Digitized Manuscripts To Europeana (DM2E) is an EU-funded Europeana satellite project. Its primary aims are to enable as many content providers as possible to get their data into Europeana and to stimulate the creation of new tools and services for reuse of Europeana Data in the Digital Humanities. Being coordinated by Humboldt-Universität zu Berlin, the duration of the project is three years from 2012 to 2015.

The DM2E model is a specialization of the Europeana Data Model[16] (EDM) for the domain of handwritten manuscripts. The EDM has been developed within the Europeana v1.0 project as an RDF-based data model for describing rich metadata records for Europeana, the European digital library. It can handle huge metadata record collections represented by heterogeneous metadata standards that must be accessible via the same platform. The EDM covers Cultural Heritage Objects (CHOs) that are collected and delivered to Europeana by diverse cultural heritage institutions. The model is as generic as possible and can be specialized for domain-specific descriptions like it is the case in DM2E.

In May the project finished the first operational version of its DM2E model (v1.0), a specialization of the EDM for handwritten manuscripts. The ontology has been developed within work package 2 with a lot of input from others in the project. Especially results of extensive mapping workshops with the data providers of DM2E were integrated into the model. Metadata of diverse formats like TEI, EAD and MARC21 was analyzed and used to create new classes and properties that specialize the current EDM.

[15] ERIC stands for European Resarch Infrastructure Consortium
[16] http://pro.europeana.eu/edm-documentation

Pundit, the semantic annotation tool developed by Italian SME Net7 as part of the DM2E project and other EC-funded projects, won the LODLAM (Linked Open Data in Galleries, Libraries, Archives and Museums) challenge in June 2013.

Principles

Linked Data is the paradigm that drives the whole DM2E infrastructure. The DM2E model reflects this by explicitly defining classes for datasets and published data resources. This way, the meta-level of resource descriptions becomes a first-class member of the data model and can be used for annotations and provenance tracking.
The specification document[17] describes the DM2E data model in its first operational version. It extends DM2E v0.2 of the DM2E project and is the current specialization of the Europeana Data Model (EDM) made by DM2E. The DM2E Model reuses as many existing properties and classes from other ontologies as possible.

DM2E enables organizations to link sections of text to each other or to other Linked Data resources on the Internet such as DBPedia, Freebase and Geonames. In case a text document comes with a microstructure including sub-entities identified by URIs such structures can be used transparently – or else a highlighting function will be available that would as well enable the highlighting of image areas.

Use cases

Pundit developed by the DM2E project partners, is a semantic annotation tool for Digital Humanities that enables scholars to annotate digitized manuscripts and is already being used to annotate the digitized manuscripts being made available to the project.

In March 2013, 5000 pages of Wittgenstein Archive were introduced into Pundit and a group of scholars is now doing humanities research with this tool. This pilot has been a great way to collect feedback from a scholarly community and further develop Pundit to the needs of humanities researchers. At present an international group of scholars is using Pundit to annotate Wittgenstein's work as part of a DM2E research experiment called the Wittgenstein Incubator.

Finally, August saw the publication of a paper written by Alois Pilcher of the University of Bergen and leader of the Wittgenstein Incubator initiative, which will see Wittgenstein scholars work with digitized Wittgenstein manuscripts held at Bergen. The paper entitled "Sharing and debating Wittgenstein by using an ontology" was published in the journal of Literary and Linguistic Computing and draws heavily on the research and work related to the DM2E project.

Moreover, University Library Johann Christian Senckenberg in Frankfurt will provide a set of medieval manuscripts using DM2E model. Adding these collections to Europeana via the DM2E project will result in richer metadata that can greatly improve the research possibilities for humanities scholars.

[17] http://dm2e.eu/document/#DM2EModelSpecification

Advantages and drawbacks

Advantages include the following:

- Adding the new collections to Europeana via DM2E will result in richer metadata that can greatly improve the possibilities for humanities scholars for research
- Pundit will make it easier for libraries, archives and museums to provide data on their collections to Europeana and avoid the time-consuming (and sometime inaccurate) job of mapping the various data formats of third-party archives to Europeana's overarching classification system

Drawbacks include the following:

- The software Pundit is currently in testing phase in several institutions
- An open issue of consistency check and update to the newest EDM version
- As a content provider, the content you deliver to DM2E will be integrated into the Europeana platform at the end of the project. For this reason, the digital objects made available by the institution (facsimiles, full-text transcriptions, etc.) need to be licensed in accordance with Europeana requirements

Europeana Data Model and Functional Requirements for Bibliographic Records

Overview

The EDM – FRBRoo Application Profile Task Force (EFAP-TF) was launched in response to a recommendation from Europeana V1.0, a project which had the core task of transforming the Europeana prototype into an operational service. This recommendation asked for an application profile that would allow a better representation of the FRBR group 1 entities: work, expression, manifestation and item.

The Final Report on EDM – FRBRoo Application Profile Task Force[18] identifies two important motivations for understanding its findings:

- The application profile is not a prescriptive framework for producing new object representation metadata within Europeana, it is not a set of cataloguing rules – instead it is strictly limited to the mapping of existing source data to a specialized EDM framework.
- The intention is to create buy-in from two communities – the Europeana community and the FRBRoo community – in order to make the connection of the two worlds as seamless as possible. This motivation had some influence on the composition of the Task Force in that there is a conscious effort to include people from the FRBRoo context.

Principles

The measurements of success for the Task Force are considered to be the timely deliverable of:

- Combined model in terms of properties and classes

[18] http://pro.europeana.eu/documents/468623/1760978/TaskfoApplication+Profile+EDM-FRBRoo

- Principles for modelling and mapping rules

The deliverable will be used by those who model derivative relations in the Europeana Data Model. The final report delivers combined models in terms of properties and classes of EDM and FRBRoo supported by three example data samples provided by the Task Force.

Linked Heritage project

Overview

Linked Heritage[19] is an EU co-funded project that aims to extend and enrich the metadata holdings of the Europeana digital library. The project builds on the earlier Athena project[20] and runs from April 2011 until 30th September of 2013. The project had tackled a number of tasks, including the coordination of terminologies, standards and technologies used, linking of cultural heritage data into the semantic web, training, and the ingestion of assets into the Europeana collection itself.

The Linked Heritage consortium, continuing the work of the earlier ATHENA and MINERVA[21] projects, has developed a well understood and tested standard method for aggregating cultural heritage data for preservation, standards development and experimentation, and contribution to Europeana.

It is important to note that the main items highlighted in Linked Heritage were about web semantic, linked data and the state of the art in cultural metadata models (in particular their interoperability across libraries, museums, archives, publishers, content industries, and the Europeana models).

This project is as a first investigation to determine the precise extent of progress in practical semantic interoperability between the whole cultural heritage and commercial sectors.

Standards for use in linked heritage

Based on the survey conducted by the project it was decided to use LIDO as the primary metadata standard for aggregation within the Linked Heritage project. LIDO (Lightweight Information Describing Objects) is the result of a collaborative effort of international stakeholders in the museum sector to create a common solution for contributing cultural heritage content to portals and other repositories of aggregated resources, as well as exposing, sharing and connecting data on the web.

Further criteria for the selection LIDO cover:

- Being built upon previous work, and the large experience of international stakeholders in the museum documentation area, LIDO gained a widespread

[19] http://www.linkedheritage.org/
[20] http://www.athenaeurope.org/
[21] http://www.minervaeurope.org/

adoption in a very short amount of time. It has established a large user base and support within the CIDOC[22] community.
- LIDO" s interoperability has been proved with metadata used by the different content providers, as well as interoperability with both Europeana" s ESE and EDM standards.
- The technical implementation of LIDO in the metadata interoperability services (MINT[23]) that will be used in the Linked Heritage project, were developed during the ATHENA project. The solution has proved successful already for the ingestion of large amounts of data into Europeana.
- The schema design process for LIDO v1.0 took into account from the beginning the requirements for implementing the linked data concept, and in particular persistent identification so it is a suitable choice for integration with linked data technologies.

For the library domain there seems to be no established ingestion workflow beyond Dublin Core / ESE data. Therefore since an important goal of the Linked Heritage project is the enrichment of Europeana, e.g. through the provision of as rich metadata as available, it will be examined what the library community is planning for future ingestion into Europeana, and a mapping template will be provided for transforming data from MARC variants used by providers in the Linked Heritage project, into LIDO.

3.2 OCR and linguistic resources

Preserving and providing OCR results enables a lot of new opportunities for a broad range of users, especially in the context of digital humanities. Additional information kept together with OCR results, such as linguistic resources, makes these possibilities even much greater. The following subsections provide an overview of the related standards and formats.

Analysed Layout and Text Object (ALTO)

Overview

ALTO was initially developed by the METAe European project group for use with the Library of Congress' Metadata Encoding and Transmission Schema (METS). While METS excels in describing the structure of objects, a schema related to the content and layout information of each piece of the object was missing.

CCS (Content Conversion Specialists GmbH) maintained the ALTO standard until August 2009, when the Library of Congress (LC) Network Development and MARC Standards Office became the official maintenance agency for the ALTO XML Schema. The ALTO Board thus oversees maintenance of the ALTO XML Schema and helps foster usage in the digital library community.

[22] http://cidoc.icom.museum
[23] http://mint.image.ece.ntua.gr/redmine/projects/mint/wiki

Principles

ALTO stores layout information and OCR recognized text of pages of any kind of printed documents like books, journals and newspapers. ALTO can detail technical metadata for describing the layout and content of physical resources (text, illustrations, graphics).

ALTO describes a content page with different views:

- The Description section helps to describe some general settings and information of the ALTO file (measurement units, file name, etc.), and the production process itself (processing steps, software used, dates and actors, etc.)
- The Layout section contains what's on the page. A page is divided into several regions (print space; left, right, top and bottom margins). For each region, all objects are listed which have been detected inside: text blocks, illustrations, graphical elements, composed blocks. Each object previously identified is defined by generic attributes: width, height, text content (for the String element). Besides, the reading order of all the elements can be managed.
- Each ALTO file may also contain a style section where different styles (for paragraphs and fonts) are listed.

Use cases

ALTO is one of the most common formats used by libraries for converting text from images. It's used both to deliver digitized contents and to preserve these contents.

In a delivery perspective, the ability of ALTO to store the text content coordinates in a page allows the overlay of image and text (multilayer PDF) and highlight search words in a query.

Figure 7 Multilayer PDF (left) and Web digital library (right) with plain text search

It most commonly serves as an extension schema used within the METS administrative metadata section, in order to preserve patrimonial contents. However, ALTO instances can also exist as a standalone document used independently of METS.

Advantages and drawbacks

ALTO takes benefits of the XML world:

- XML is readable and understandable, even by novices, and no more difficult to code than HTML.
- ALTO schema is quite simple, and therefore, ALTO contents are easily understandable.
- XML is completely interoperable: any application that can process XML can use your information, regardless of platform.
- ALTO contents can be distributed between libraries, they are interoperable, etc.
- XML contents are transformable: ALTO contents can be transformable into simple text files, HTML pages, etc.

ALTO also inherits disadvantages of XML:

- Each XML language needs adequate processing applications to display, transform contents, etc.
- ALTO needs specific tools (e.g. an ALTO file can't be displayed in a web browser)
- XML is extendable: ALTO XML schema can be hacked locally (e.g. ALTO BnF)

Besides, ALTO has shown some other limitations:

- Physical description: the layout region types supported by ALTO are limited. One may want to be more precise: maths content, music score, etc.
- Logical description: ALTO format captures the layout and the full text of OCRed pages. But one may want to mark the logical structure of documents. This can be done with a container format like METS in association with ALTO (to capture the intellectual structure of the document), and/or with logical labelling of structural elements in ALTO (page numbers, margin note, etc.)

These limitations will be addressed by the next version of the ALTO format, which is planned to be published in January 2014.

Page Analysis and Ground-Truth Elements (PAGE)

Overview

Page Analysis and Ground-Truth Elements (PAGE) is a format framework related to production and evaluation of Optical Character Recognition and Document Image Analysis results. One of the main design goals was to enable "a highly detailed and accurate description of any information which can be derived from a given document image" (S. Pletschacher, 2010) overcoming limitations of existing formats (like ALTO) and allowing its use in applications requiring a very precise content representation (such as performance evaluation). PAGE is based on a number of XML-Schemas which specify a root structure and individual sub-formats. All Schemas are maintained by the PRImA Research Lab and are publicly available at http://schema.primaresearch.org/PAGE/.

There are numerous software tools which support PAGE natively and it is the format of choice of various (large scale) reference datasets in the digital library and document analysis research community.

Structure and sub-formats

The PAGE format framework specifies a root structure to link more specific sub-formats (called gts – ground truth and storage) related to individual processing steps in a document recognition workflow (see Figure 8).

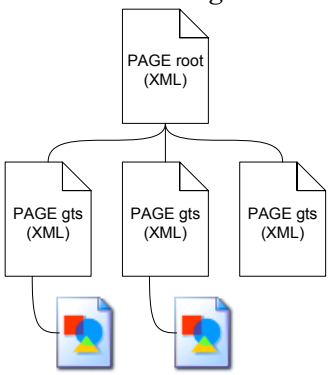

Figure 8: General PAGE structure (S. Pletschacher, 2010)

There are currently four sub-formats for Binarisation, Deskew, Dewarping, and Page Content. Page Content is by far the most widely used PAGE sub-format and therefore often referred to as PAGE format (although PAGE is the whole framework). It allows for a very accurate description of layout elements (regions with precise polygonal outlines), text (Unicode encoded), reading order (including groups of ordered and/or unordered objects), and layers (objects on different levels - like stamps on top of text regions). Text regions can be further structured into lines, words and glyphs (each allowing for full Unicode text) and may be assigned labels such as heading, paragraph, caption, page number etc. Other types of regions are image, line drawing, graphic, table, chart, separator, maths, noise and frame. Depending on the region type there are further sub-types describing the function (like stamp, handwritten annotation etc.) as well as detailed metadata (such as language, script, font, reading direction, text color, background color).

Besides the content- and processing-related sub-formats there are also formats foreseen for storing results and settings (profiles defining penalties and error weights) related to performance evaluation.

Tools and support

PAGE is supported by a number of tools, which are actively being developed and maintained by the PRImA Research Lab (www.primaresearch.org/tools). The most prominent example is Aletheia, a comprehensive ground truth production solution.

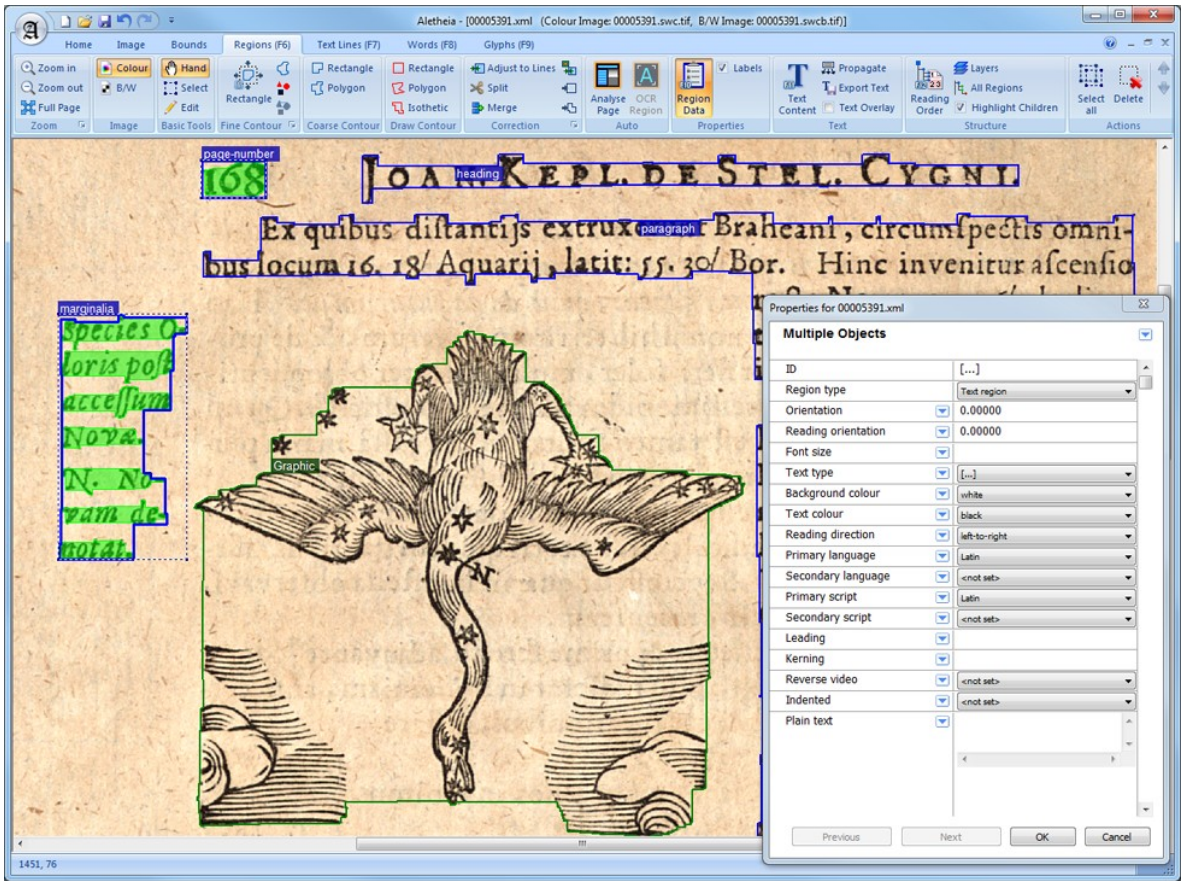

Figure 9: Aletheia – a comprehensive ground truth production tool natively supporting PAGE

Other tools include quality assurance, performance evaluation for layout analysis and OCR, OCR exporters (e.g. from FineReader Engine and Tesseract) and format converters, interactive viewers (SVG) for embedding in web-based digital libraries and repositories, as well as APIs (C++ and Java) in order to support developers implementing PAGE in third party software.

Succeed is supported by the European Union under FP7-ICT and coordinated by Universidad de Alicante.

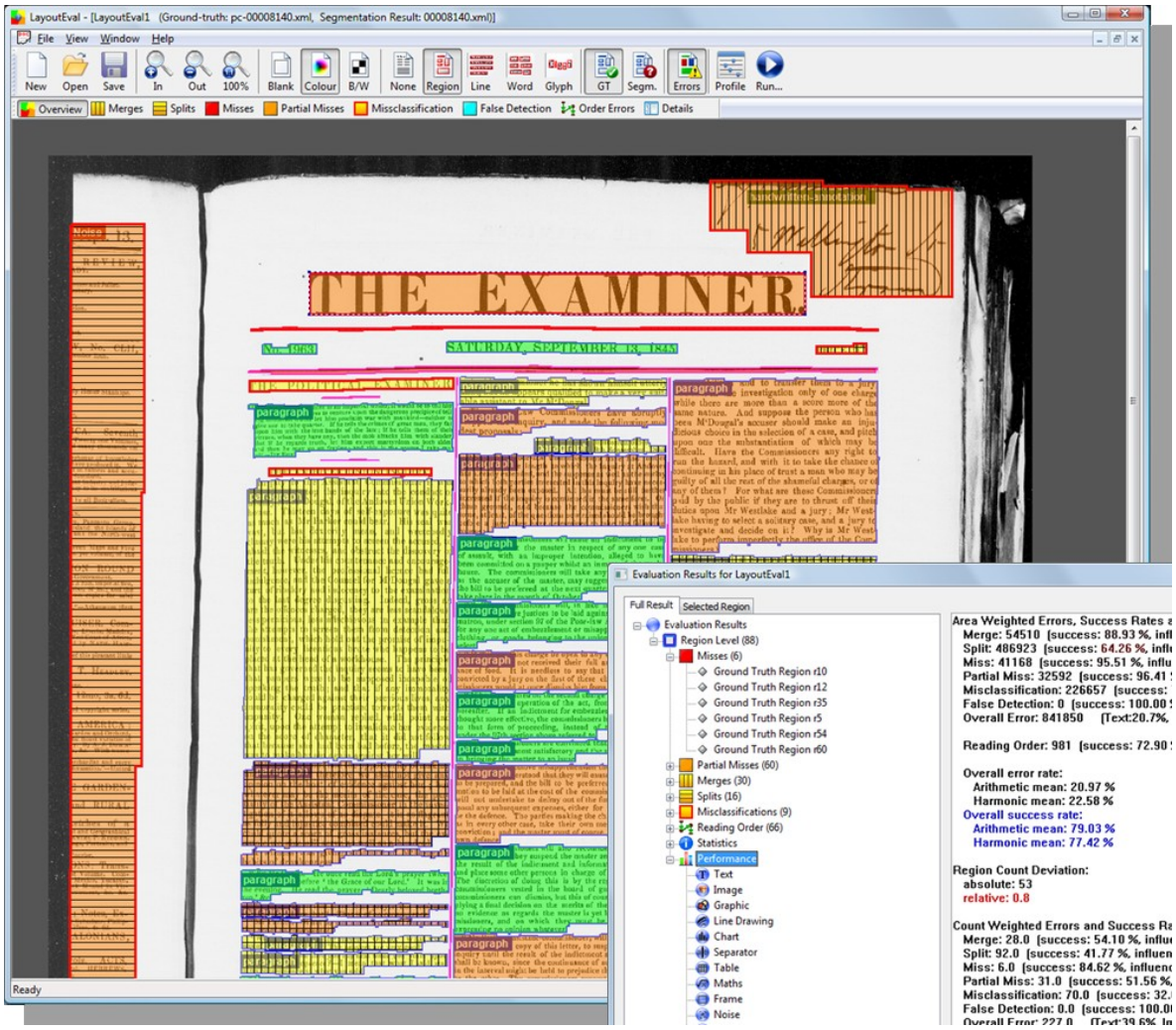

Figure 10: Layout Evaluation Tool – natively supporting PAGE

Usage

PAGE (Page Content) is used as the main ground truth format in several datasets and related large scale evaluation activities. The IMPACT Image and Ground Truth Repository (now maintained by the Impact Centre of Competence[24]) is probably the biggest of its kind with close to 50,000 manually created high-quality ground truth files (including detailed region outlines, layout information, full Unicode-encoded text – to the level of ligatures and special characters, and reading order) of historical documents. Another example is the PRImA Layout Analysis Dataset for contemporary documents[25]. Besides existing datasets, PAGE is also being used in currently ongoing digitisation activities and research projects (for instance Europeana Newspapers[26], eMOP[27],

[24] http://www.digitisation.eu/
[25] http://dataset.primaresearch.org/
[26] http://www.europeana-newspapers.eu/

tranScriptorium[28]). PAGE has also a long standing tradition as the format of the ICDAR (International Conference on Document Analysis and Recognition) page segmentation competitions (A. Antonacopoulos, 2013).

Europeana Newspapers project

Overview

Europeana Newspapers[29] is an EU ICT-PSP project with the main goal to make available 18 million pages of digitized European newspapers via a shared service of Europeana/The European Library (TEL). The project is creating full-text for 10 million pages of digitized newspapers from 12 libraries across Europe, and also developing an interface to allow for cross searching of over 18 million newspaper pages.

Principles

Within Europeana Newspapers, a metadata profile that can serve as best practice for an information package of refined digital newspapers is being developed. A main design consideration was to provide a format that allows for human readability and machine-readability at the same time. One of the challenges lies in the fact that for the first time Europeana will not only gather metadata, but actually receive information packages containing metadata, images and full-text. Therefore it was important to design an information package that provides an effective and simple solution so that the needs of Europeana can be served in an optimal way.

The suggested information package shall conform to the OAIS[30] (Open Archival Information System) standard and will be implemented as a METS[31] (Metadata Encoding and Transmission Standard) container. Data stemming from OCR (Optical Character Recognition) processes will be stored within ALTO[32] (Analyzed Layout and Textual Object) files. The working name for our information package is therefore ENMAP which stands for: **E**uropeana **N**ewspaper **M**ETS **A**LTO **P**rofile.

ENMAP has been discussed at several occasions by the Europeana Newspaper consortium. External expertise was gathered, e.g. METS profiles from the British Library and the Australian National Library were studied and evaluated for their usability within the project. ENMAP is intended to provide a simple but effective encoding for all newspapers that are refined within the Europeana Newspaper Project (ENP). A first (internal) release of ENMAP took place in September 2012, followed by an internal feedback cycle. The first public release (towards the end of 2013) will provide the suggestion for an SIP (Submission Information Package) for Europeana that shall

[27] http://idhmc.tamu.edu/emop/
[28] http://transcriptorium.eu/
[29] http://www.europeana-newspapers.eu/
[30] http://public.ccsds.org/publications/archive/650x0m2.pdf
[31] http://www.loc.gov/standards/mets/
[32] http://www.loc.gov/standards/alto/

also create the main prerequisites for successful preservation actions. Nevertheless the question of born-digital newspapers will not be covered by this format.

Advantages and drawbacks

The advantages of ENMAP cover:

- ENMAP is based on practical considerations drawn from the experience of refining 10 million newspaper pages from 12 European libraries.
- ENMAP clusters different kinds of information about a digital newspaper in standardized container formats.
- ENMAP stores information in a way that make it easy to use for various purposes (be it refinement, online presentation or preservation).

Drawbacks include:

- ENMAP may be too comprehensive for smaller institutions who do not have a large newspaper collection.
- There may not be many digitisation service providers that can deliver ENMAP without additional cost.

Implementation

The guidelines are set up and maintained by the Europeana Newspapers project. The Europeana Newspapers project is partially funded under the ICT Policy Support Programme (ICT PSP, http://ec.europa.eu/ict_psp) as part of the Competitiveness and Innovation Framework Programme by the European Community.

10 million pages of digital newspapers will be produced in the ENMAP format by the partners of the Europeana newspapers project.[33]

Text Encoding Initiative

Overview

The TEI (Text Encoding Initiative) is an international organization founded in 1987 to develop guidelines for encoding machine-readable texts in the humanities and social sciences. 'TEI' is also used to refer to the TEI Guidelines themselves, and to the set of schemas they describe.

The TEI Guidelines for Electronic Text Encoding and Interchange define and document a markup language for representing the structural, renditional and conceptual features of texts. These guidelines are expressed as a modular, extensible XML schema, accompanied by detailed documentation, and are published under an open-source license.

[33] http://www.europeana-newspapers.eu/consortium/project-partners/

Since 1994, the TEI Guidelines have been widely used by libraries, museums, publishers, and individual scholars to present texts for online research, teaching and preservation.

Since 2001, the TEI is organized as a member-funded consortium. Its chief deliverable is a set of Guidelines, which specify encoding methods for machine-readable texts.
The most recent version of the TEI guidelines is known as P5. It was initially released in November 2007 and has been updated since then on a six-month cycle.

Rules and recommendations made in these Guidelines are expressed in terms of what is currently the most widely used markup language for digital resources of all kinds: the Extensible Markup Language (XML), as defined by the World Wide Web Consortium's XML Recommendation. However, the TEI encoding scheme itself does not depend on this language; it was originally formulated in terms of SGML (the ISO Standard Generalized Markup Language), a predecessor of XML, and may in future years be re-expressed in other ways as the field of markup develops and matures.

TEI customisation
The TEI provides a number of basic, general-purpose customizations. One of the best-known of these is **TEI Lite.**

TEI Lite has been widely adopted, particularly by beginners and by big institutional projects that rely on large teams of encoders to markup their documents.

Although there is no default schema, TEI P5 does provide a number of example customizations which may very well meet your needs, which can be downloaded from the TEI web site or from within the Roma interface (see Table 18).

Table 18 TEI customisations provided by the TEI consortium

Name	Description
TEI Lite	TEI Lite, the most widely used TEI customization; includes basic elements for simple documents
TEI Tite	A constrained customization designed for use by keyboarding vendors.
Bare	TEI Absolutely Bare, a very barebones schema with the absolute minimum of elements
Corpus	TEI for Linguistic Corpora, includes the modules for encoding linguistic corpora
MS	TEI for Manuscript Description, includes the elements for describing manuscripts and complex physical aspects of documents
Drama	TEI with Drama, includes the TEI drama module
Speech	TEI for Speech Representation, includes the TEI module for spoken language
Odds	TEI for authoring ODD, includes the TEI module for creating ODD files and customizations
allPlus	TEI with with all modules included, plus all external additions
TEI + SVG	TEI with SVG
TEI + Math	TEI with MathML

TEI + XInclude | TEI with XInclude

Usage

The recommended way to customize the TEI is to create a formal specification expressing your customizations, as an XML document using TEI ODD markup; this can then be compiled into a suitable DTD, RELAX NG schema or W3C Schema (together with the appropriate reference documentation), using the Roma[34] program. Advanced users can also create the ODD by hand using normal XML editing tools.
If, however, you intend to make extensive use of the TEI in conjunction with other schemas written in RELAX NG, working directly with the RELAX NG modules is probably the best skill to learn. Typical TEI users are more likely to work solely within the confines of the TEI, and may need to use DTDs or W3C Schema as well as RELAX NG, and so writing customizations in the TEI's own language is usually better.
There are several important reasons why this high-level method is recommended:

- It is independent of the schema type (DTD, RELAX NG schema, W3C schema) and the resulting specification can be used to generate a schema in any of these schema languages.
- It lets you document your work using the familiar TEI markup.
- It provides full access to the TEI class system.
- The Roma utilities generate a single, portable, schema file which you can transfer to other people without worrying about link dependencies.

Lexical Markup Framework

Objectives

The goals of LMF are to provide a common model for the creation and use of lexical resources[35], to manage the exchange of data between and among these resources, and to enable the merging of large number of individual electronic resources to form extensive global electronic resources.

Types of individual instantiations of LMF can include monolingual, bilingual or multilingual lexical resources. The same specifications are to be used for both small and large lexicons, for both simple and complex lexicons, for both written and spoken lexical representations.

Description

LMF is one of the members of the ISO/TC37 family of standards. LMF is composed of the following components:

- The core package that is the structural skeleton which describes the basic hierarchy of information in a lexical entry.
- Extensions of the core package which are expressed in a framework that describes the reuse of the core components in conjunction with the additional

[34] http://www.tei-c.org/Roma/
[35] http://en.wikipedia.org/wiki/Lexical_resource

components required for a specific lexical resource.

The core package is depicted in Figure 11.

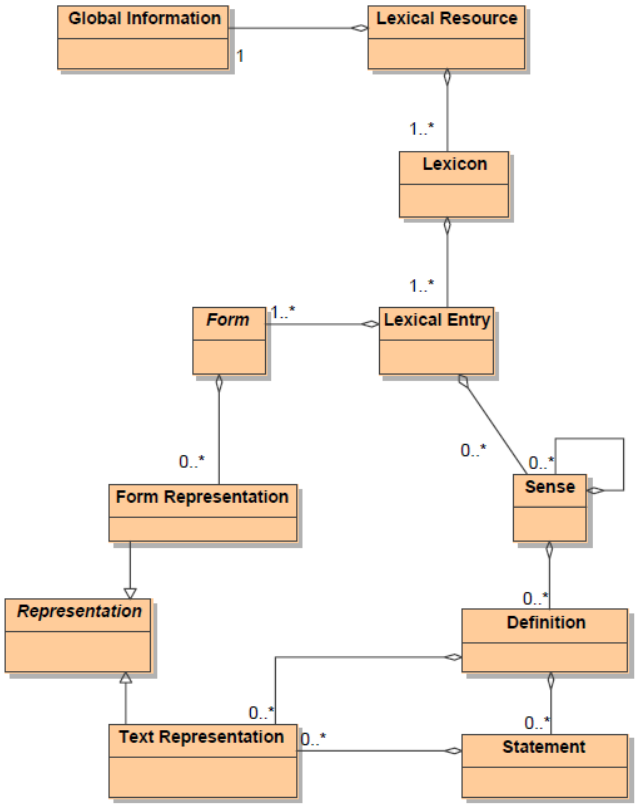

Figure 11: Core Package of LMF

The extensions are specifically dedicated to morphology, machine readable dictionary, syntax, semantics, multilingual notations, morphological patterns, multiword expression patterns, and constraint expression patterns.

Additional extensions can be defined by users. These should adhere to a number of key standards. Data elements should be organized in classes and subclasses and the data elements should also be registered as data categories at ISOcat[36].

The Open Language Archives Community

Description

OLAC, the Open Language Archives Community, is an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources by:

[36] http://www.isocat.org

- Developing consensus on best current practice for the digital archiving of language resources, and
- Developing a network of interoperating repositories and services for housing and accessing such resources.

Standards for metadata

The OLAC metadata set is based on the Dublin Core (DC) metadata set and uses all fifteen elements defined in that standard. To provide greater precision in resource description, OLAC follows the DC recommendation for qualifying elements by means of element refinements or encoding schemes.

The qualifiers recommended by DC are applicable across a wide range of resources. However, the language resource community has a number of resource description requirements that are not met by these general standards. In order to meet these needs, members of OLAC have developed community-specific qualifiers, and the community at large has adopted some of them as recommended best practice for language resource description. These recommended qualifiers are listed in OLAC-Extensions[37].

The XML implementation of OLAC metadata follows the "Guidelines for implementing Dublin Core in XML".

Medieval Unicode Font Initiative

Overview

The Medieval Unicode Font Initiative[38] is an international workgroup with the goal of harmonizing and eventually standardizing the encoding and display of special characters, which can be found in historical documents but are not currently part of Unicode.

Goals

Unicode is today's commonly recognized standard for encoding documents which may be potentially multilingual and based on multiple scripts. As far as modern languages and writing systems are concerned Unicode covers a large number of alphabets and special characters. For historical material, however, this support is not quite as complete. The lack of standardized code points for such characters is a major impediment for digitization projects dealing with historical documents. MUFI is therefore pursuing two major goals: As a short term solution, unprecedented characters are collected, clarified, and assigned a code point in the Private Use Area of Unicode. An important aspect of this is to coordinate the allocation of code points and to maintain the character recommendation list. As a long term solution, MUFI is proposing such characters to Unicode in order for them to be included in future versions of the standard.

[37] http://www.language-archives.org/REC/olac-extensions.html
[38] http://www.mufi.info/

Character recommendations

MUFI is maintaining a list of current character recommendations which are agreed by the workgroup and coordinated with some other initiatives like the TITUS project and Junicode. Following these recommendations for characters which are not part of Unicode (yet) will minimize the risk of code point collisions (different meaning for the same code point) between different digitization projects and software solutions. The MUFI character recommendations are provided in alphabetical as well as code chart order[39].

Font pages

Closely related to the maintenance of recommendations is the problem of how to display documents using this intermediate character encoding solution. MUFI is therefore also working on fonts which provide graphical representations for the above mentioned Private Use Area code points as defined in the recommendations[40].

3.3 Relevant ERICs

ERICs play an important role in Europe's innovation, especially in the context of joint establishment and operation of research infrastructures. In the following subsections practices of two digitization-related ERICs are presented.

Digital Research Infrastructure for the Arts and Humanities (DARIAH)

The Digital Research Infrastructure for the Arts and Humanities (DARIAH) aims to facilitate long-term access to, and use of, all European Arts and Humanities (A+H) digital research data. The main goals of this initiative is to create an infrastructure of people, information, tools, and methodologies for investigating, exploring and supporting work across the broad spectrum of the digital humanities. This vision also includes the exploration and definition of common standards to ensure interoperability of metadata and data across different locations, different disciplines, different scholarly and cultural traditions as well as different languages.

To foster the interoperability of tools in the A+H domain, which may include digital objects like text, music, images or other artifacts, DARIAH relies on the widely adopted Dublin Core standard in combination with the Text Encoding Initiative (TEI) format. Since simple Dublin Core is usually not sufficient to give an extensive description of digital objects, domain-specific profiles such as CEI (charters), epiDoc (epigraphic data) or MEI (music) are used to provide a more fine-grained description.

The DARIAH project contributes to the establishment of common standards, comparable to TEI, in various research areas to foster the interoperability between the numerous institutions. Additionally, DARIAH may also give impulses to the further development of TEI such as the treatment of Hebraic texts in TEI (currently developed by the Research Data group) or the development of a persistent identifier (PID) meta-resolver. DARIAH

[39] http://www.mufi.info/specs/
[40] http://www.mufi.info/fonts/

will also organize expert seminars in order to gain feedback from humanities scholars concerning the ways in which they use TEI tools, the needs they have for the further development of them, with a specific focus on how they wish them to interact with TEI markup.

Common Language Resources and Technology Infrastructure (CLARIN)

Overview

CLARIN is the short name for the Common Language Resources and Technology Infrastructure, which aims at providing easy and sustainable access for scholars in the humanities and social sciences to digital language data (in written, spoken, video or multimodal form) and advanced tools to discover, explore, exploit, annotate, analyze or combine them, independent of where they are located (see http://clarin.eu/).

Metadata principles

CLARIN endorses two methods to describe metadata: CMDI, ISOcat. These will be described below.

CLARIN proposes the CLARIN Metadata Initiative (CMDI), a component-based approach: you can combine several metadata components (sets of metadata elements) into a self-defined scheme that suits your particular needs: a profile. Profiles and components are stored in a Component Registry to promote reuse.
There are tools available to convert other existing metadata formats like Dublin Core and OLAC to CMDI.

The data categories mentioned in the metadata (CMDI) should be registered at ISOcat[41]. ISO 12620 provides a framework for defining data categories (DC's) compliant with the ISO/IEC 11179 family of standards. According to this model, each DC is assigned a unique administrative identifier, together with information on the status or decision-making process associated with the DC. In addition, DC specifications in the Data Category Registry (DCR) contain linguistic descriptions, such as DC definitions, statements of associated value domains, and examples. DC specifications can be associated with a variety of data element names and with language-specific versions of definitions, names, value domains and other attributes.

3.4 Application packaging

Introduction

Software maintenance is one of the most crucial parts of the software development life-cycle. Because it is a post-deployment activity it brings a lot of potential risks, including loss of applicability, increased complexity or lower quality. Specific challenges that need to be faced in this context have been identified by Lehman and stated as a series of laws

[41] http://www.isocat.org

of software evolution[42]. Some of the key aspects of software maintenance include simple, efficient and cost-aware procedures for installation, upgrades, patches and retirement. Such activities can be usually performed in an automated manner (supervised by the system administrator). Recommended way for such activities is application packaging, which means creating operating system aware packages for a particular software tool. These packages have a potential to substantially reduce the cost of maintenance, and in the context of open-source tools have a potential to increase their take up thanks to simplified (in many cases automatic) installation procedure.

Packaging for Operating Systems

Operating systems (OS) determine how the application needs to be installed in order to work properly. Prescription includes the rules for storing executable files, defining dependencies as well as documentation. Multiple tools exist to support building and verifying application packages for all the leading operating systems.

In general there are two types of packages: binary and source. The difference is that the binary package is installed as a ready to use executable, while the source package needs to be compiled before installation. Source packages are mostly used in Linux-based OS and MacOS (e.g. by using MarPorts). In case of Windows it is more common to work with binary packages or even ready to run software executables.

Tools that support creating packages are usually based on scripts and files structure (including descriptive files about version, dependencies, etc.) that should be installed into the target OS. For example Debian-based Linux OS uses .deb packages, which can include various information like metadata about the software, necessary libraries, executable files, shell script to run, etc. The final package can be provided as a single .deb file, which can be easily installed, using the OS build-in package management tool. The .deb package is composed of multiple items, but the most important elements are:

- The control file which contains information about the software, including dependencies, version, maintainer, short description, install size, etc.
- The copyright file, which contains intellectual property information, especially license under which the software is distributed.
- The changelog file consisting of information about all the changes that has been performed in the upstream code.
- The rules file stating how the files need to be handled during the build procedure in order to actually build a package.
- Documentation files, which include manpages, examples, extended documentation, etc.

There are multiple application packaging tools available for different OS. Most common examples, which mostly originate from the open-source community are:

- In case of Debian-based Linux OS: dh tools, lintian, dpkg-deb.

[42] http://users.ece.utexas.edu/~perry/work/papers/feast1.pdf

- In case of MS Windows examples are WiX Toolset, MAKEMSI , NSIS.
- In case of MacOS this can be freely available PackageMaker.

Packaging for cloud technologies

Cloud technologies are currently an important element of the IT infrastructures. More and more services are offered via cloud technologies, which usually bring reduction of costs thanks to optimized consumption of storage and processing resources. In the context of digitization it is already visible and present on the market. The examples include DuraCloud, Omeka.net or ABBYY Cloud solution.

Cloud computing is highly based on a virtualization, which enables cost effective management of available hardware components. Virtual servers are usually configured using the virtual machine image, which makes it possible to create or run a new virtual servers. The examples cover different formats, including OVF (Open Virtualization Format), Amazon Machine Images (AMI), Virtual Disk Image (VDI), QEMU copy-on-write version 2 (qcow2). Additionally to that there are initiatives, such as Application Packaging Standard (APS), which provide standard ways of software packaging for cloud Software As A Service (SaaS) model. It means that having the tool packaged according to APS it is possible to easily deploy it on any SaaS cloud, which supports the standard.

Digitization and tools packaging

Digitization process is usually composed of many steps that are performed manually or automatically (e.g. scanning can be done manually, while conversion to delivery format can be automated). In order to follow mass digitization practices it is important to enhance digitization workflow with as many automated processing as possible. This is usually done by integration of new tools and resources into the digitization workflow. For the simplicity of the integration process, wider take-up of tools and ease of their further maintenance it is important to use packaging techniques. There are already available good examples of such an approach, e.g. software packages maintained by Open Planets Foundation[43] or IMPACT Center of Competence[44]. Such packages can be directly used to deploy or update particular tools on the targeted operating system.

3.5 Summary of ongoing and emerging activities

Recent activities in the context of semantic technologies indicate that it is an important aspect for the whole information society. More and more projects and institutions aim at leveraging semantic technologies and opening up their data, by making them available according to the Linked Open Data paradigm, The Europeana Data Model (EDM) is one of the most prominent examples that make it possible to share data in a more informative way than commonly used in Dublin Core based metadata harvesting. There are multiple activities around EDM, including those, which aim at increasing its interoperability, e.g. EDM – FRBRoo application profile task force. Another example is

[43] https://bintray.com/openplanets
[44] https://bintray.com/impactocr

the D2ME project which builds on EDM and therefore makes sure that content delivered to Europeana by project partners reaches a high level of information and interoperability, which is especially valuable in the context of digital humanities (e.g. semantic annotations). In the Linked Heritage project the LIDO format has been selected as a way of providing digital content to Europeana. One of the criteria, which indicated LIDO, was interoperability with EDM. In the context of this project it is interesting to note that LIDO seems to be a good solution for quite extensive description of digital objects, which covers the needs of various cultural heritage institutions. At the same time it provides a good opportunity to expose data with the open data paradigm in mind.

OCR representation and supporting linguistic resources seems to be an important element in every digitization project. Preservation of this information is important from the perspective of the user (e.g. more possibilities for automated analysis) as well as the content holder (e.g. lowering costs by OCR results reuse when producing new delivery formats). Current awareness of these needs is limited, as only 11 recommendations and practices (described in section 2) tackled this issue in their documents. The most common format for OCR results representation is ALTO. It is widely used by various institutions across the world, including recent activities such as the Europeana Newspaper project. The ALTO format has limitations, which mostly origin from the fact that it was designed as a supplement for the METS format. Therefore specific functions are assumed to be handled by the METS and not the ALTO format itself (e.g. logical structure). The PAGE (Page Analysis and Ground-Truth Elements) is a XML-based format framework related to production and evaluation of Optical Character Recognition and Document Image Analysis results. One of its goals was to overcome limitations of the existing formats (like ALTO). Although the format addresses various needs of OCR results representation it is not widely used by the cultural heritage community. Another possibility for the OCR results representation are TEI guidelines which provide principles for encoding machine-readable texts useful for humanities and social sciences. The coverage of the TEI format is extensive, including various features of texts. There are multiple customizations of TEI, including TEI for linguistic corpora and TEI lite for simple documents. A dedicated format for representing linguistic resources is LMF, which is a part of ISO/TC37 family of standards. The OLAC metadata set is developed by the institutions, which create virtual library of language resources. The metadata set is an extension of the Dublin Core and provides a common way of describing language resources, such as corpuses, transcriptions, word lists, etc.

ERICs relevant in the context of textual resources and digitization investigate various approaches for improved research. DARIAH is leveraging TEI and Dublin Core standards for describing digital objects, and contributes to comparable standards as well. CLARIN is focused on language resources and proposes a CLARIN Metadata Initiative (CMDI) for describing them.

The last aspect of the digitization process investigated in this section is application packaging, which is highly important element, especially in the context of small and medium cultural heritage institutions, where IT expertise is not really advanced.

Particular operating systems have their own packaging techniques and tools for creating such packages are available and ready to use. Also cloud technologies increase their presence in digitization field, therefore packaging formats and standards are also important aspect to investigate in this context.

4. SUCCEED SURVEY ON FORMATS AND STANDARDS

In order to provide an overview of the digitization-related standards and formats used by cultural heritage institutions across Europe it was decided to develop and conduct an online survey. The following sections elaborate on the purpose and scope of the survey, methodology as well as results analysis.

4.1 Purpose and scope

The purpose of the survey was to gather information about current practices from various cultural heritage institutions in the context of text digitization. The overview of the results was an input material for Succeed project recommendations. It was assumed that participating institutions are active players in the context of digitization, which means that they execute digitization related activities, such as digitization projects, digital library maintenance or data preservation.

The survey itself consists of 30 questions divided into 5 sections. The sections are focused on:

- General information such as contact information, institution type, etc.
- Long-term preservation aspects including questions related to used metadata and data formats as well as experience in preservation and digitisation.
- Online delivery of digital assets, including metadata and data formats with OCR aspects in mind.
- Emerging standards, formats and approaches to discover innovative activities in the context of OCR and digitisation workflow enhancement.
- Standards related to digitisation tools including questions related to used operating systems and tools packaging.

The survey questionnaire is available in the The original survey filled in by respondents is available at: http://bit.ly/succeed_wp4.

4.2 Methodology

The survey was prepared in form of an online questionnaire. The questionnaire has been prepared in a series of consultations with Succeed project partners, based on their experience with digitization, including tools, content creation, preservation, analysis and online delivery. There were two types of questions:

- Option questions – a question consisting of several options to mark (one or many), including an "Other" option to provide a response, which does not appear on the options list.

- Open questions – question consisting of an input field where respondent can answer with free text.

The option questions were used when there was a set of most probable options to be pointed by respondents. An example is a question about master files where TIFF format is one of the options. Open questions were used in cases where there is no clear answer to the question, e.g. preferences in the context of tools packaging.

It was decided to create an online survey to reach a wider community and simplify the procedure of answering to the survey. For efficient and successful dissemination of the survey a list of dissemination channels has been created. It was composed of two main parts:

- List of institutions to directly ask to fill in the survey – it includes 31 institutions to which Succeed partners have direct contacts and can with high probability obtain answers to the survey.
- Other dissemination channels, such as mailing lists, blogs, etc. – a list of 15 channels (hundreds of institutions) to which information about the survey should be sent.

Each item from those two lists has an indication of responsible partner. It means that indicated partner was responsible for dissemination of the Succeed survey via this particular channel. Dissemination activities were done in two rounds, each of them lasted for approximately one week. In each round all partners from Succeed project were asked to disseminate information about the survey to the channels they were assigned. First round of dissemination gave approx. 60 responses. The second dissemination round increased the number of responses by approx. 20. The final number of responses is 86 and those answers were further analyzed.

4.3 Analysis of results

Survey analysis is divided into several sections, each presenting different topics of investigation. It is important to mention that for many questions in the survey respondents could give more than one answer (e.g. more than one master file format could be indicated), therefore sums (or percentages) of responses related to particular options in a certain questions can be higher that the number of responses (or higher than 100%).

Overview

Altogether 86 respondents from different countries across the world have filled in the survey. The respondents were from various institutions, including libraries, archives, museums, research institutes and companies. The respondents were mostly from Europe, but there were also responses from North America, South America, Africa, Asia and Oceania. Figure 12 presents the percentage of the respondents coming from different continents across world. The other series includes international and institutions identified with less than 100% accuracy (e.g. wrong e-mail provided).

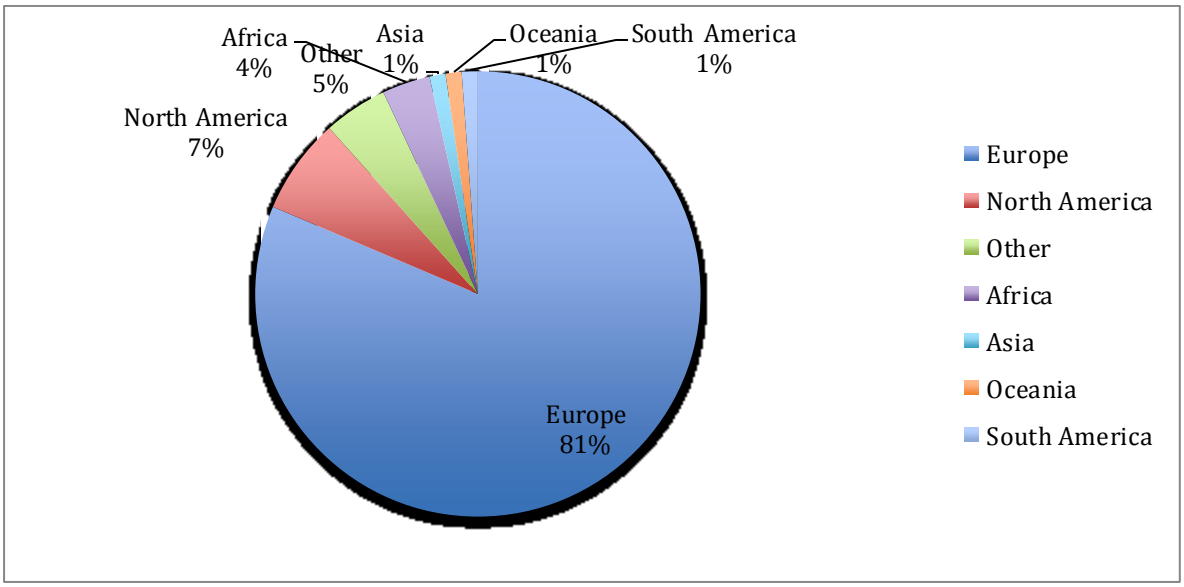

Figure 12 Respondents of the survey divided by continent

Europe itself has been covered quite well both in terms of number of responses (70) as well as geographical coverage (see Figure 13) in terms of countries from which institutions provided feedback.

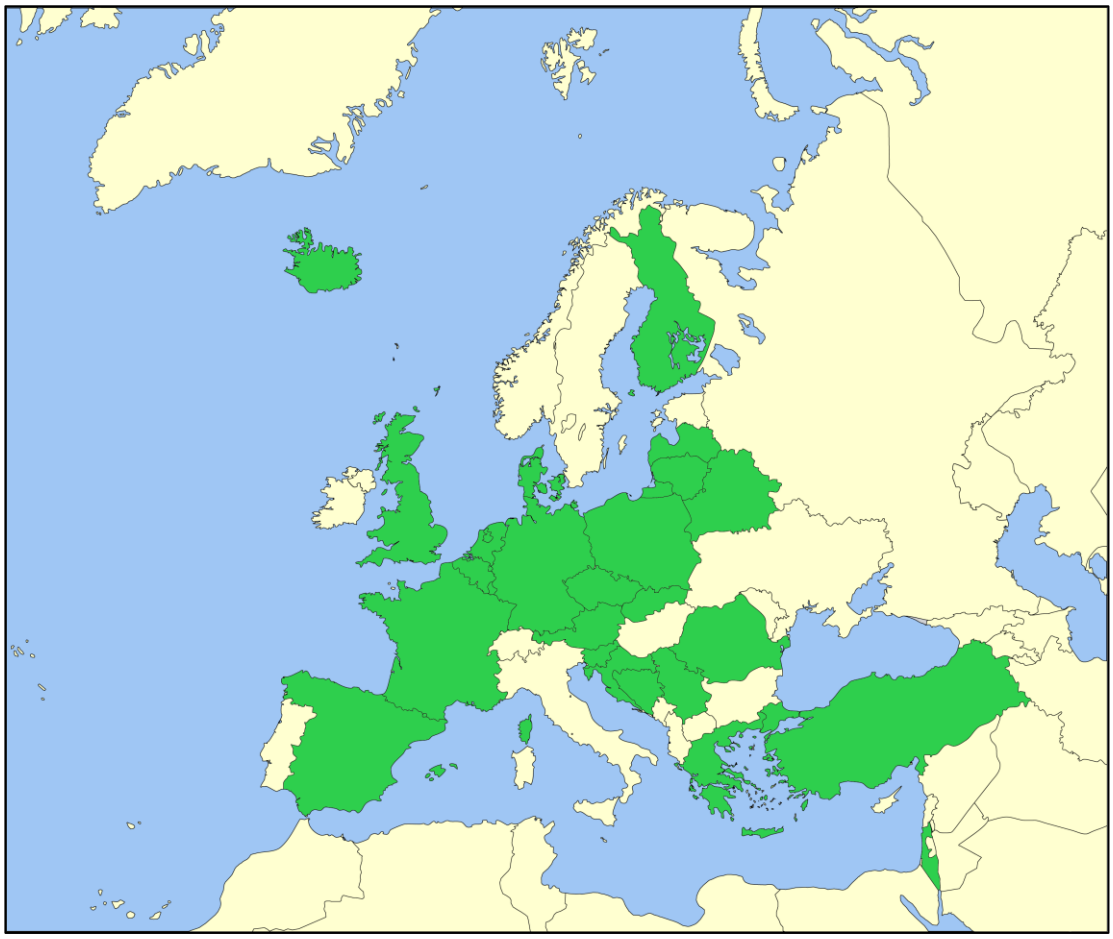

Figure 13 European countries covered by the survey (green marked areas indicate European countries from which institutions participated in the survey)

Long-term preservation

The long-term preservation section of the survey investigated file and metadata formats used for preservation. The responses were also analyzed in the context of preservation experience and digitization experience. Preservation experience has been divided into four groups:

- Very large – in case of institutions that have already preserved more than 1 000 000 of digital pages
- Large – in case of institutions that have already preserved less than 1 000 000 pages but more than 100 000 pages
- Medium – in case of institutions that have already preserved less than 100 000 pages but more than 10 000 pages
- Small – in case of institutions that have already preserved less than 10 000 pages

And digitization experience has been divided into five groups:

- Very large – in case of institutions that in the last 5 years digitise per year more than 250 000 pages

- Large – in case of institutions that in the last 5 years digitise per year more than 100 000 but less than 250 000 pages
- Medium – in case of institutions that in the last 5 years digitise per year more than 10 000 but less than 100 000 pages
- Small – in case of institutions that in the last 5 years digitise per year more than 1 000 but less than 10 000 pages
- Very small – in case of institutions that in the last 5 years digitise per year less than 1 000 pages

Master file formats analysis (see Figure 14) indicates that TIFF format is the most popular across all respondents, getting almost 90% of indications. There is no other format that is equally popular. All other indications gain less than 25%. From the other master formats PDF and JPEG2000 are the options to consider in the context of still images. Although JPEG has been indicated as well by 23% of respondents it does not seems to be a good option to consider due to its lossy compression. Technical characteristics for the master files were indicated in the context of used master file format. The most common format is TIFF, and its characteristics are as follows:

- DPI: at least 300, depending on the document type. It has been noted several times that quality index[45] should be at the level of 8 (excellent quality in the context of letters size in the document)
- Colour depth: at least 24-bit for colour images and 8-bit for greyscale images
- Compression: Most of the respondents use uncompressed images (64%), if compression is used then LZW is mostly used

In case of JPEG2000 the most common compression filter is 5-3 reversible.

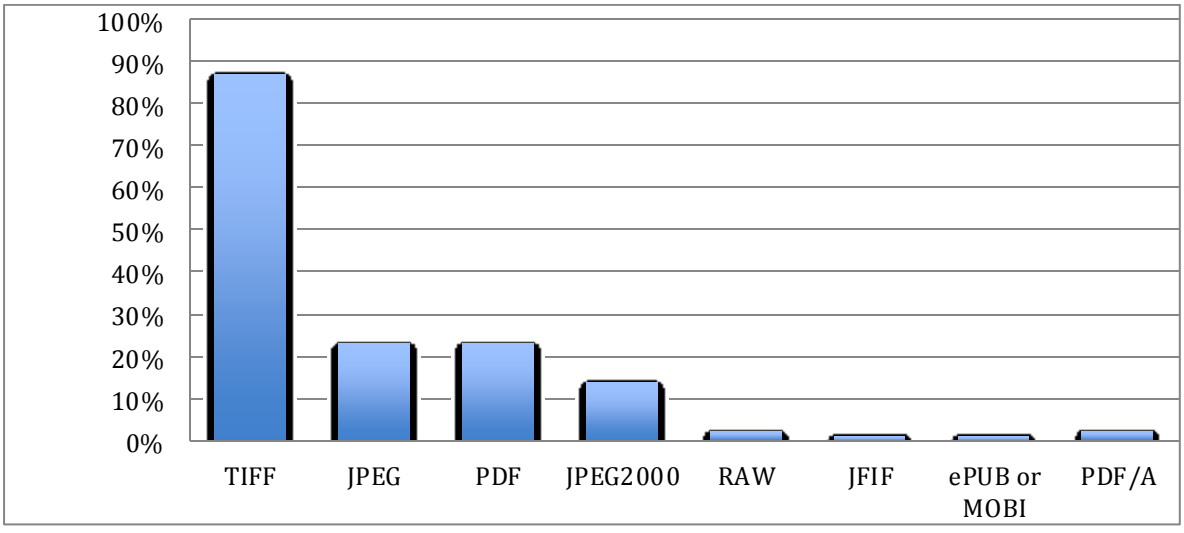

Figure 14 Master file formats used in the context of long term preservation

[45] http://www.clir.org/pubs/abstract//reports/pub53

The usage of master file formats in the context of institutions preservation experience is presented on the Figure 15. Four most popular formats have been analyzed in this chart. There are no strict dependencies, except that in case of TIFF it is visible that the larger the institution the more probable it will use TIFF as the master file. PDF and JPEG usage is to some extent converged, while JPEG2000 is not really commonly used.

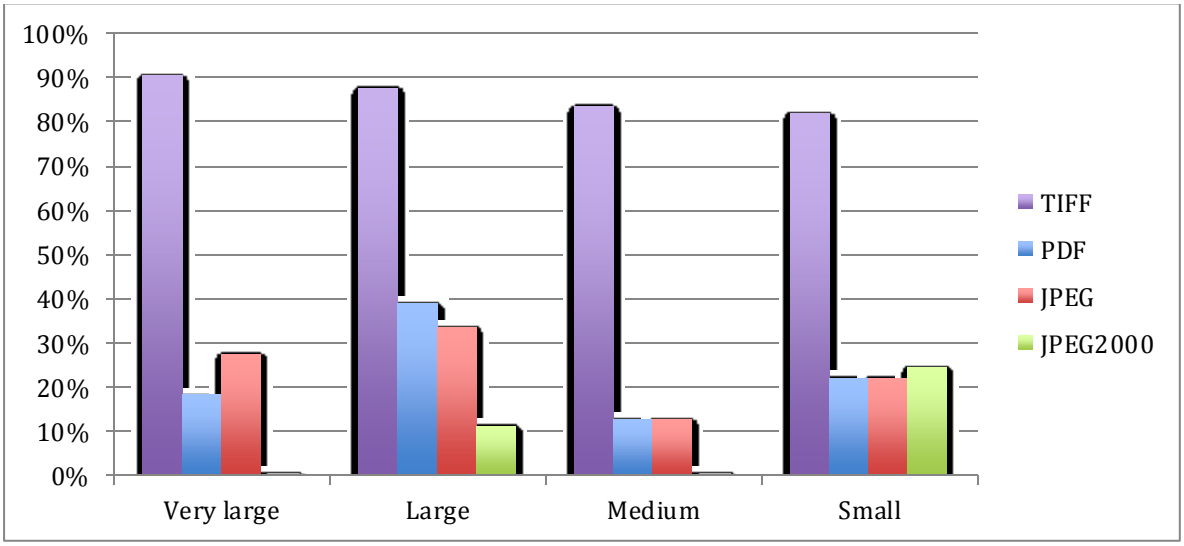

Figure 15 Usage of master file formats in the context of institution's preservation experience

Usage of master files in the context of digitization experience (see Figure 16) shows that again, with higher institutions experience comes higher probability that the institution will use TIFF as a master file.

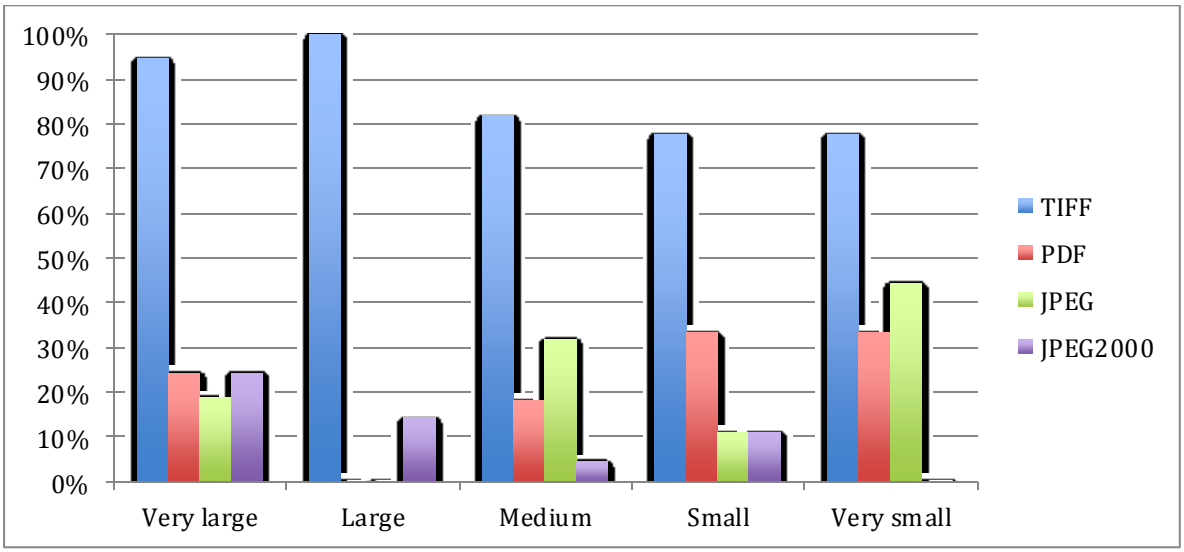

Figure 16 Usage of master file formats in the context of institution's digitisation experience

In the context of metadata formats there is a large number of indicated options (see Figure 17). The most popular formats for descriptive metadata include Dublin Core, MARC and MODS. For the preservation metadata it is most common to use PREMIS, while for the structural metadata it is usually METS. Technical metadata are usually encoded using NISO Z39-87 data dictionary.

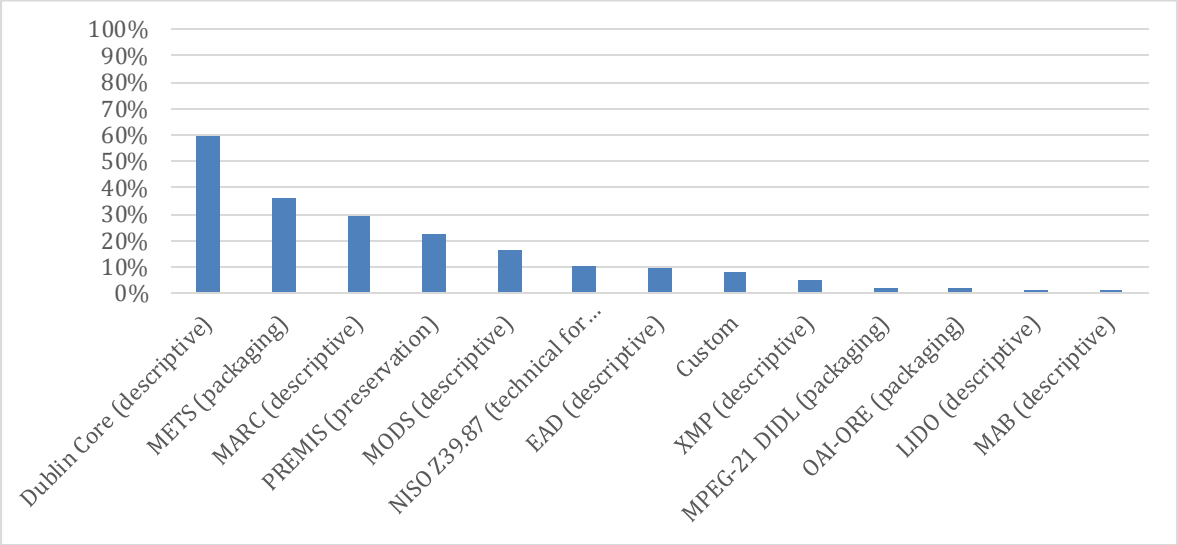

Figure 17 Metadata formats used in the context of long term preservation

When preserving OCR results (see Figure 18) the most common approach is to store PDF file (51%). There are also practices related to storing OCR results using XML with coordinates (28%) and plain text (27%).

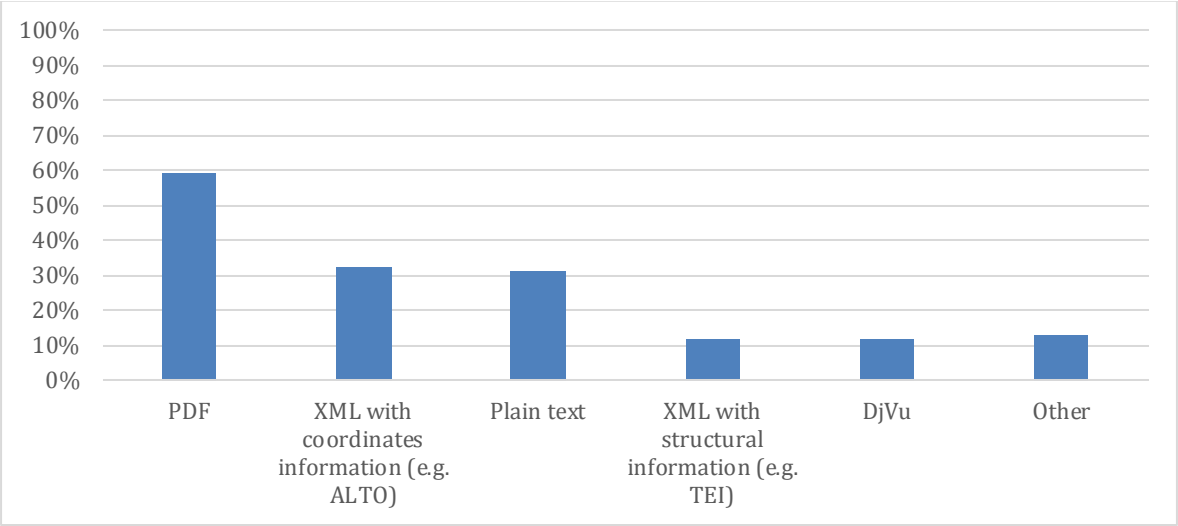

Figure 18 File formats used in preservation of OCR results

From the analysis of the responses of the question about online availability of master files, it is visible, that the majority of institutions (more than 70%) do not want to make their master files available online (see Figure 19). Within this group only 9% cannot make them available due to the copyright issues. Nevertheless, almost 30% of respondents would make their collections of master files available.

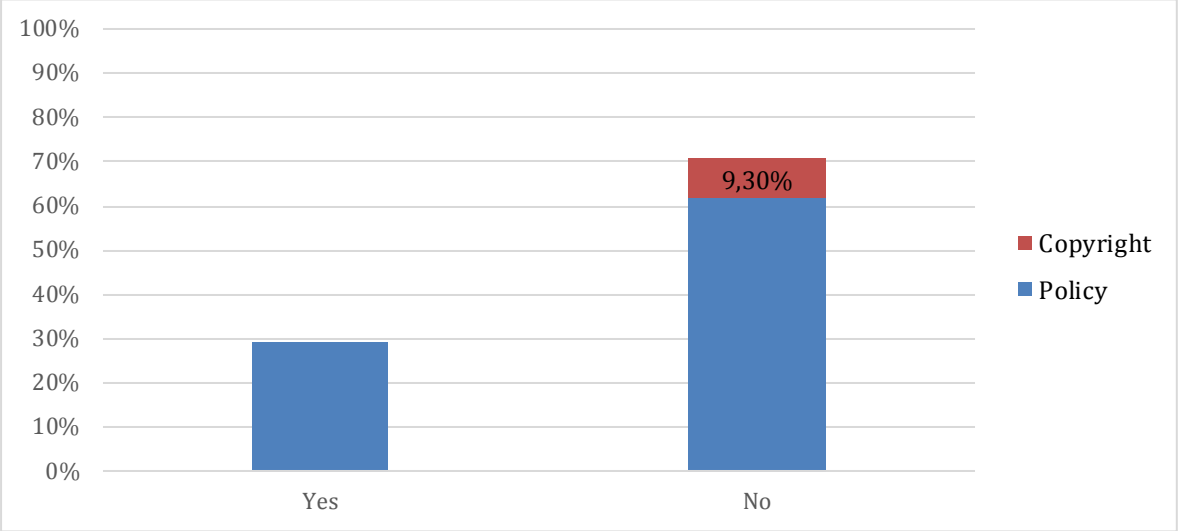

Figure 19 Willingness to make the master files available online

Online delivery of digital objects

The online delivery section of the survey investigated file and metadata formats as well as OCR results format used to provide online access to delivery files.

In case of delivery file formats the most popular are PDF (76%) and JPEG (70%). The other formats include DjVu (very popular in Poland), JPEG2000 and PNG. See Figure 20 for a summary or responses.

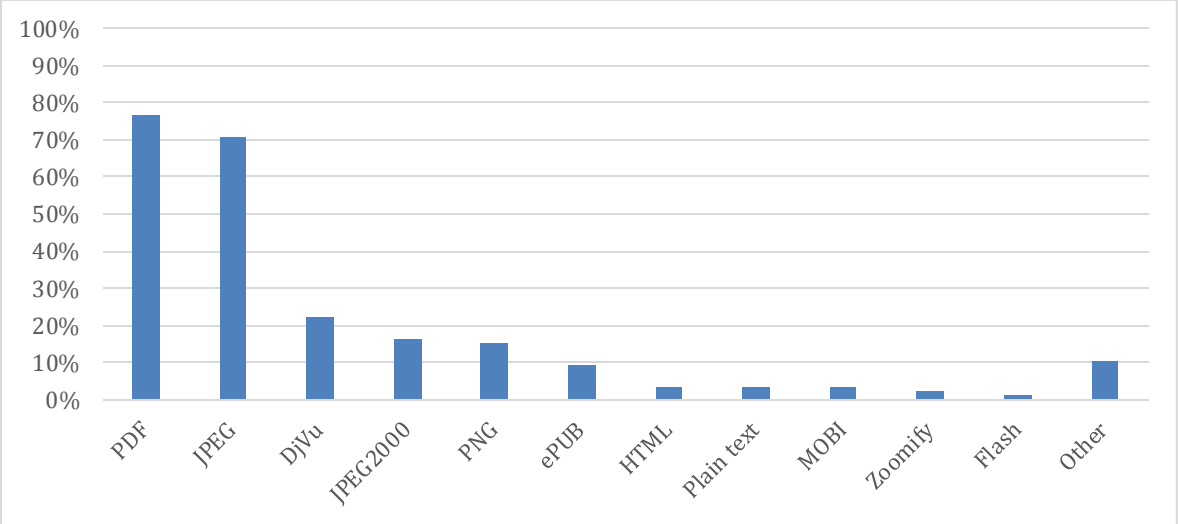

Figure 20 File formats used to provide online access to digital objects

Metadata formats provided online are available mostly in Dublin Core format (almost 70%). There is a strong user group providing metadata using Europeana Data Model (more than 20%) and Europeana Semantic Elements (more than 10%). It is therefore visible that Europeana is an important service for content providers, because it is especially supported by more than 30% of respondents. It is supported despite the fact, that Europeana can integrate with services having only Dublin Core support. Please see Figure 21 for a summary.

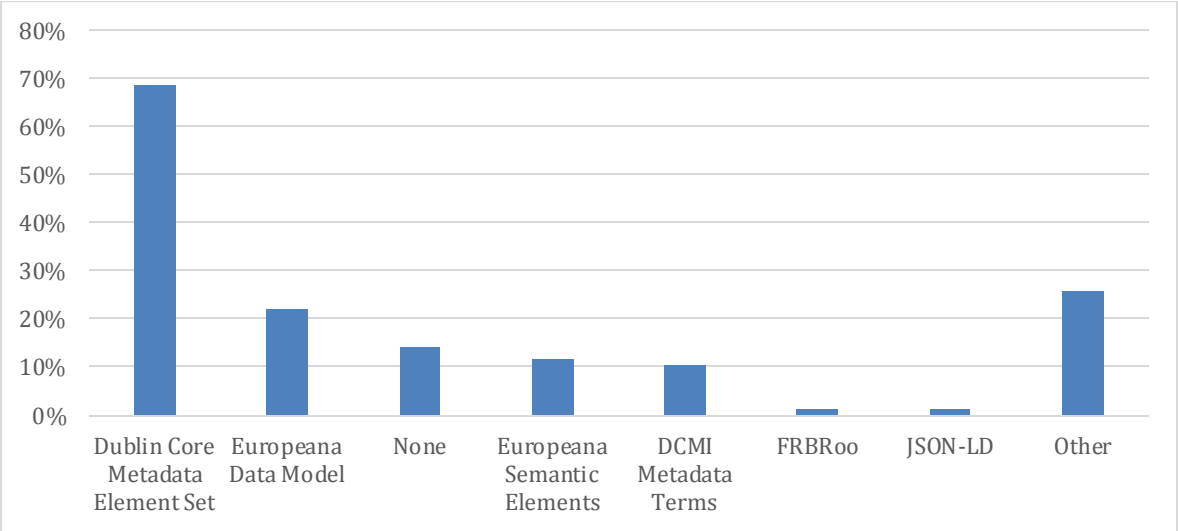

Figure 21 Metadata formats provided for the online users or aggregation services

Access to OCR is usually provided via integrating it with delivery file, e.g. PDF with text layer or an image with highlighted text. Nevertheless, still more than 25% of respondents do not provide OCR results at all. The summary of responses is available on Figure 22.

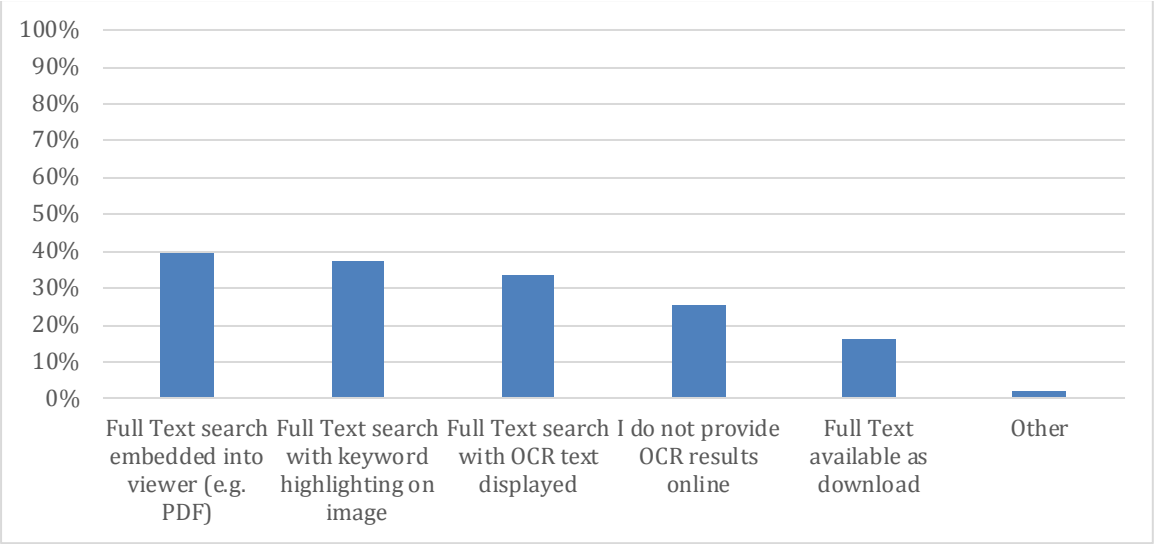

Figure 22 Approach for providing full text search and access to OCR

Emerging standards, formats and approaches

This section of the survey investigated Linked Open Data usages among the respondents, as well as usage of other advanced technologies to enhance digital content. This part of the survey investigated also OCR evaluation methods and linguistic resources formats to get a general understanding of advances and innovative technologies being currently in use by cultural heritage institutions.

The usage of Linked Open Data (LOD) to provide digital content is largely limited (see Figure 23). More than 66% of respondents do not provide digital content with the use of LOD paradigm in mind. If used then RDFa and SPARQL are pointed as semantic technologies used to give access to digital content. The usage of semantic technologies is more visible in institutions with larger experience (number of pages online).

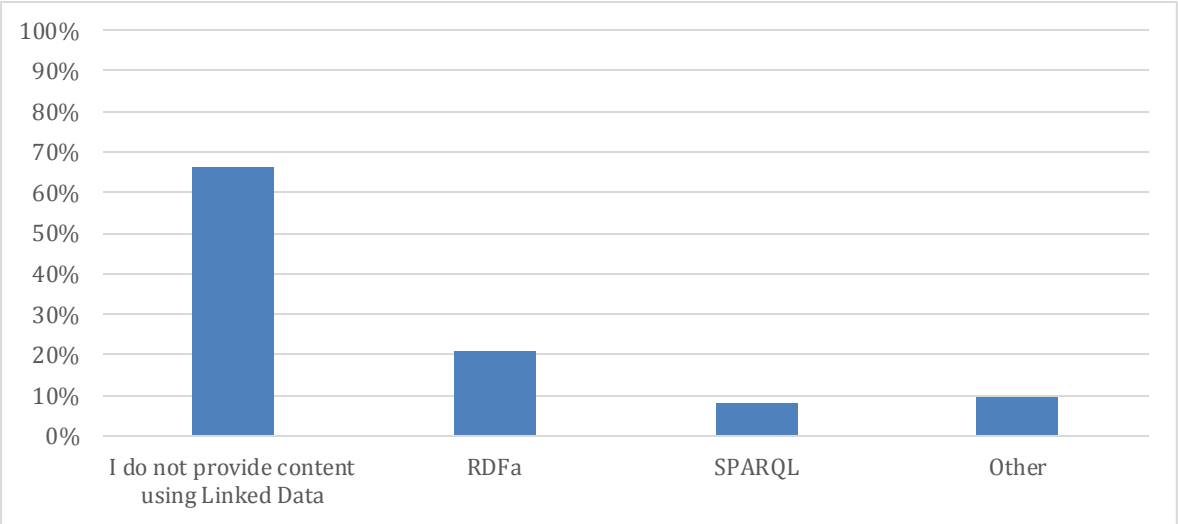

Figure 23 Usage of Linked Open Data to provide digital content

The usage of persistent identifiers for digital objects is presented on the Figure 24. The OAI Identifies is the most common selection, while Handle System and DOI are the second one.

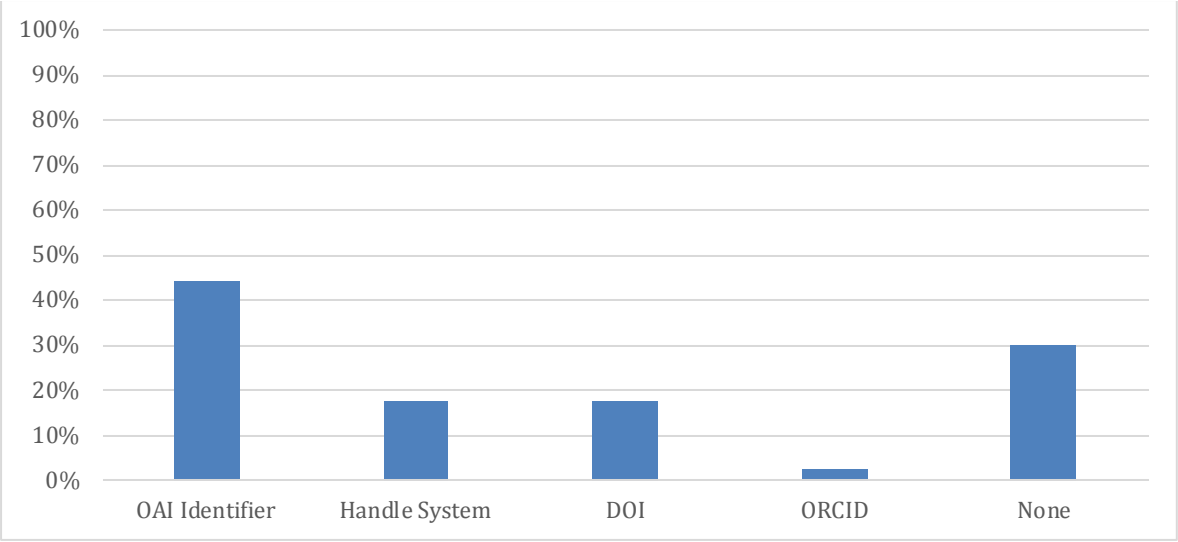

Figure 24 Usage of persistent identifiers for digital objects

The advanced technologies were indicated by a very small group of respondents – approx. 6% (for a summary see Figure 25). In this group the following technologies have been indicated:

- Named Entities Recognition Engine
- Geolocation
- OCR correction
- Automatic translation

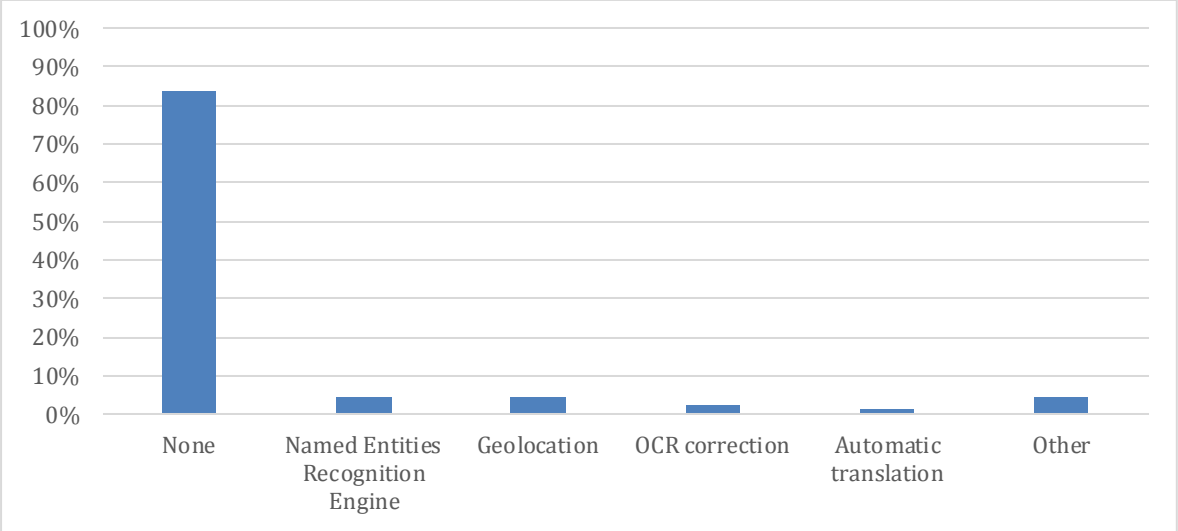

Figure 25 Usage of advanced technologies to enhance/enrich digital content

Similarly to advanced technologies, the usage of formats to store linguistic resources is very limited (see Figure 26). The formats indicated by respondents who store linguistic resources include TEI, CMDI and LMF.

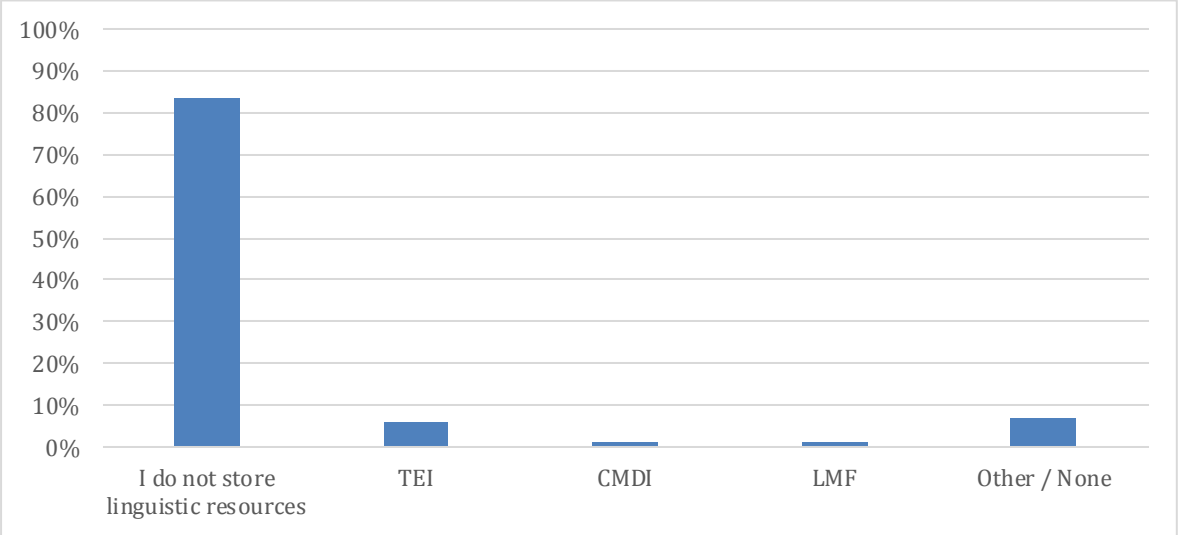

Figure 26 Usage of formats to store linguistic resources

In case of OCR evaluation two aspects have been investigated, namely OCR text recognition and OCR layout recognition. The results are similar in both cases (please see Figure 27 and Figure 28). Most of respondents (approx. 60%) do not check the quality, and if yes the evaluation is either based on OCR engine statistics or dedicated evaluation activities. Quality checks are more popular in the context of text recognition (41% of respondents do it) than in the context of layout recognition (34% of respondents do it).

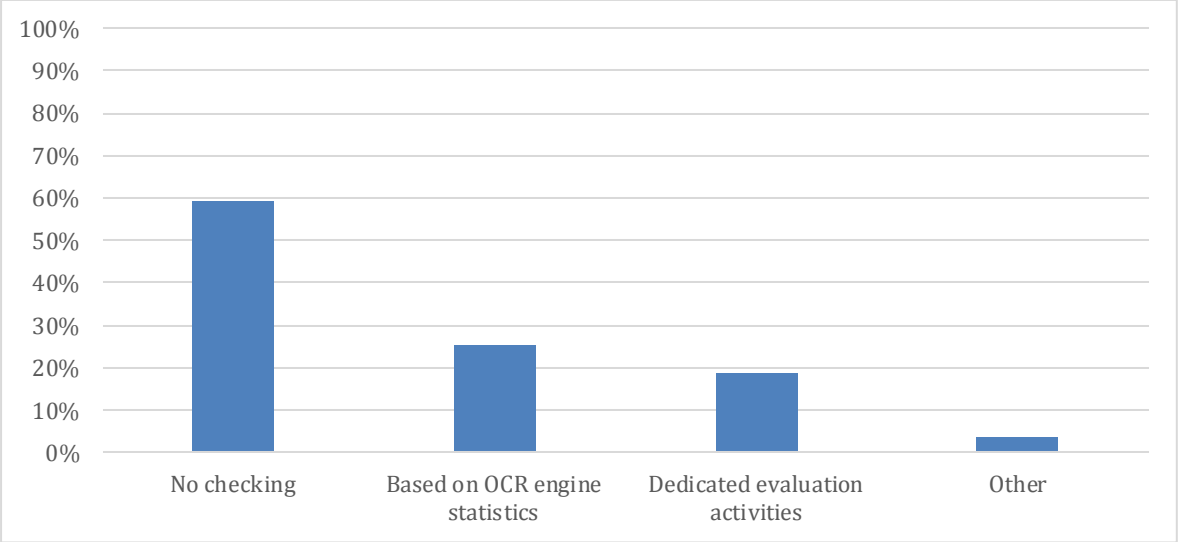

Figure 27 Usage of evaluation methods for OCR results

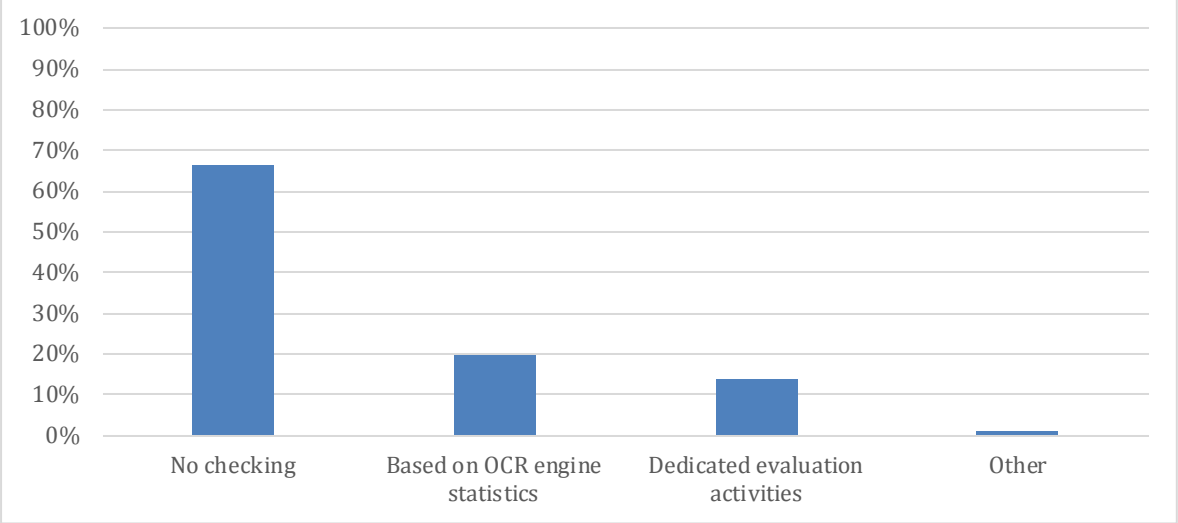

Figure 28 Usage of OCR evaluation methods for layout recognition

Standards in digitisation related tools

Tools play a critical role in digitization activities. Tools provide functionality to organize work, perform conversions, execute OCR, etc. The standards in digitization related tools section investigated working environment of the responding institutions.

The Operating System (OS) is a key component in the software infrastructure. The majority (87%) of respondents indicated MS Windows as the working OS. Linux systems on the other hand are used by almost 50% of respondents. Each of Unix and MacOS systems is used by approx. 10% of respondents. A summary is available on Figure 29.

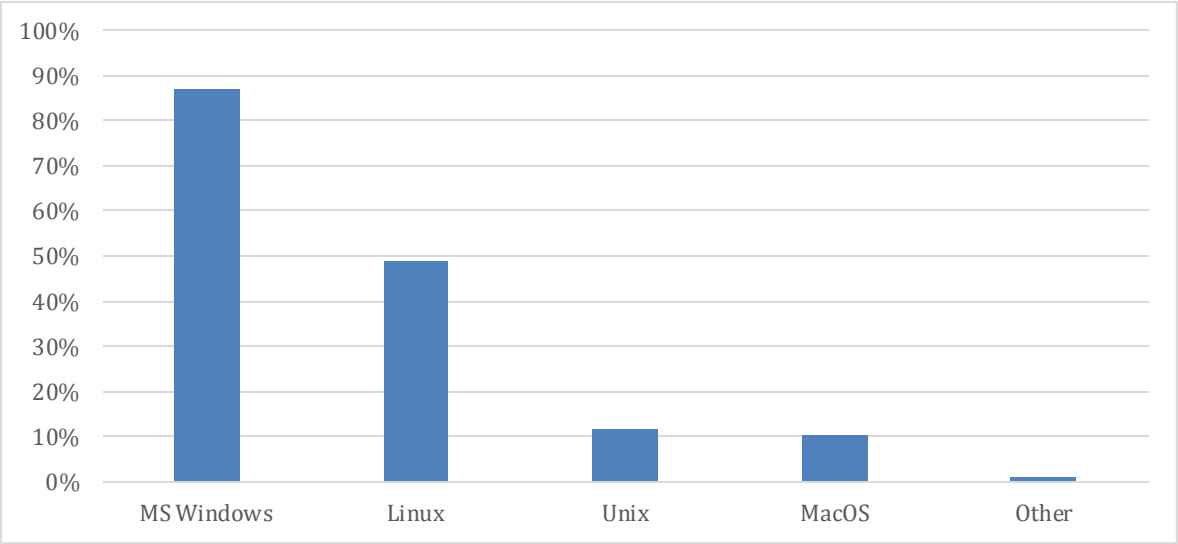

Figure 29 Usage of Operating Systems at cultural heritage institutions

Digitization workflow management systems are not largely popular among respondents (see Figure 30). Almost 55% of respondents do not use digitization workflow management system at all. There is no most popular system among the respondents, as the "Other" option has been indicated by most of those using such a system. Indicated systems include docWorks, dLab (used mostly in Poland) and Goobi (used mostly in Germany).

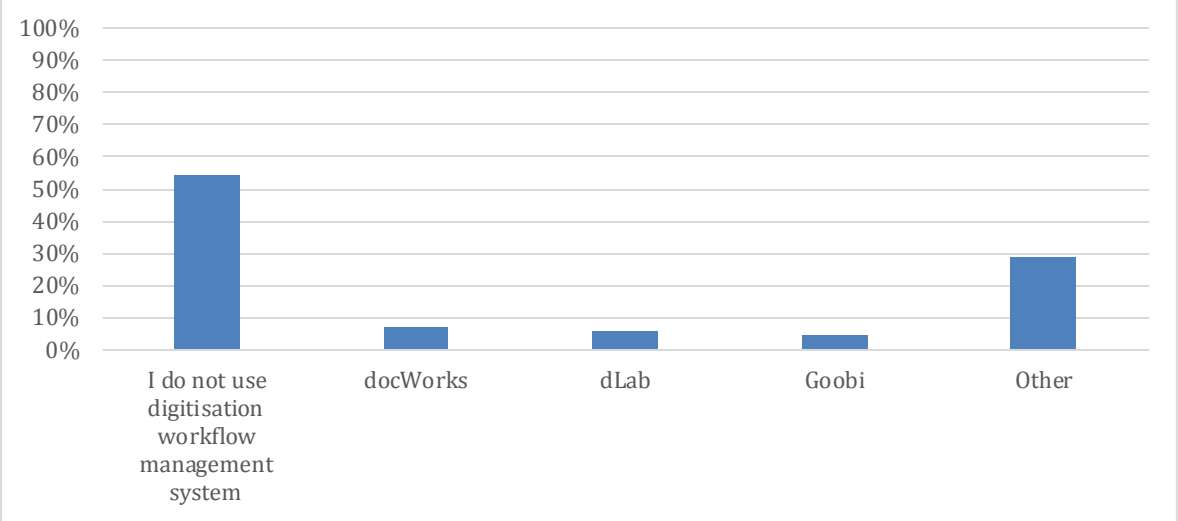

Figure 30 Usage of digitisation workflow management systems

The situation is different (in comparison to digitization workflow management system) when it comes to digital library system for online delivery of digital objects. There is a wide range of systems which make it possible to provide digital content online, including

those very popular in particular countries (like dLibra[46] system in Poland), and those used worldwide (like DSpace or EPrints). From the summary (see Figure 31) it is visible that all of respondents use such a system, but most of them use other than listed on the Figure 31.

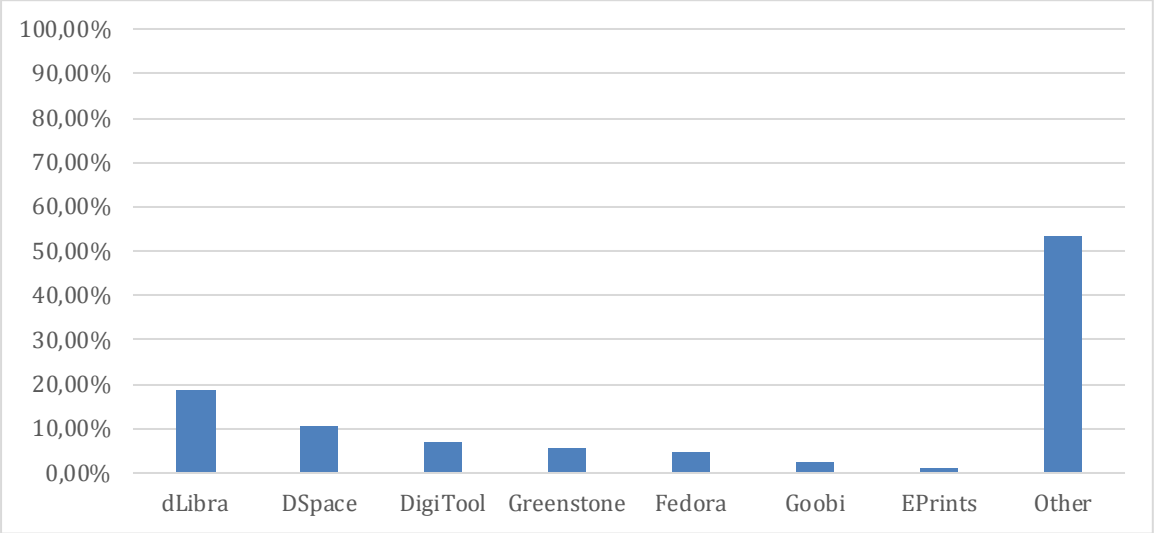

Figure 31 Usage of digital library systems for online delivery of digital objects

5. SUCCEED RECOMMENDATIONS

This section provides a set of recommendations for enhanced interoperability and preservation of the text/printed materials. The aim of these recommendations is to help stakeholders (research groups, companies and cultural heritage organizations) to select a particular format or standard for their digitization-related activities. The recommendations are divided into 3 parts, each focused on a specific aspect of digitization activities:

- Long-term preservation – this part covers formats and standards related to master files, metadata and OCR results.
- Online delivery – this part covers formats and standards related to delivery files, descriptive metadata, OCR results and identifiers.
- Advanced and supporting technologies – this part covers guidelines for semantic technologies, linguistic resources and tools packaging.

The above division is dictated by practical reasons – if particular institution performs digitization for preservation activities, then attention should be put on the long term preservation part. If the institution wants to perform digitization for access, then online delivery part should be of interest. If the institution does both (digitization for preservation and access) then long term preservation and online delivery parts should be

[46] http://dingo.psnc.pl/dlibra

investigated. Finally, if there is a vision of using new and advanced technologies to enhance the digitization workflow, then part related to advanced and supporting technologies is relevant for considerations.

Each particular aspect discussed in this section can have several recommended and alternative items (e.g. formats or standards). For instance, "Master file format – textual documents" part has TEI and PDF/A as recommended formats and UTF-8 encoded plain text as an alternative. It means that TEI and PDF/A are equally applicable and can be selected based on specific preferences or experience of particular institution. It also means that UTF-8 has some limitations which caused it to be an alternative, but not the first selection. Nevertheless if the institution does not have appropriate resources to create PDF/A or TEI documents (e.g. no appropriate software or lack of staff) or have other reason for not using the recommended items (e.g. policy), then the alternative format is proposed and can be considered as good. In discussed example UTF-8 is an alternative format and it will in most cases require a lot less effort to create. But even if an institution decides to use an alternative format, it should look for opportunities to move to the recommended one, as it is the most appropriate way of dealing with particular digitization aspect.

5.1 Long-term preservation

This part of recommendations covers formats for master files, descriptive metadata, structural metadata, administrative metadata and OCR results. The reason for selecting particular format as a recommended one is strongly connected with its sustainability factors[47], especially disclosure and adoption.

Master file format – still images

Recommended: TIFF
Alternative: JPEG2000 (JP2)

For preservation of still images the recommended format is TIFF. It is the most popular format both in the context of existing recommendations (94% of them indicate TIFF) and Succeed survey results (87% of respondents indicated TIFF). The format is well documented and has strong support in software related to scanning, OCR, manipulation and conversion. The recommended characteristics of the TIFF format are presented on the Table 19.

Table 19 Summary of recommended characteristics for the TIFF format

Characteristic	Recommendation
Spatial resolution	At least 300dpi. The final resolution should depend on the document type. The goal is to have all important characteristics of the document clearly visible. Quality Index[48] can be helpful when calculating final

[47] http://www.digitalpreservation.gov/formats/sustain/sustain.shtml
[48] http://www.clir.org/pubs/abstract//reports/pub53

	resolution.
Colour depth	24-bit for colour images, 8-bit for greyscale
Compression	Uncompressed or LZW compression
Version	6.0
Byte order	Little endian
Colour profile	ICC-based[49]
Number of pages	1 per file (monopage TIFF)

The alternative master file format is JPEG2000 Part 1 (Core) – JP2. The format is quite popular in existing recommendations (53%), but not so much in use in current digitization activities (14% of respondents of the Succeed survey use it for master files). It seems that in terms of format usage JPEG2000 looks like an emerging format rather than a well-established one. The format is well documented, but is also quite complicated. It has the capability to act both as a master file and delivery file, therefore it is especially interesting to consider for production master files. Unfortunately JPEG2000 does not have wide support in terms of software, although there are ongoing activities that develop tools supporting JPEG2000 in various ways (e.g. OpenJPEG[50], Jpylyzer[51], IIIF[52]). Because of these current limitations it has been identified as an alternative format.

Master file format – textual documents

Recommended: TEI, PDF/A
Alternative: UTF-8 encoded plain text

For preservation of documents available in textual form we recommend using of TEI or PDF/A.

TEI is focused on texts representation, including various characteristics like structural or conceptual. The format is very flexible which can be both advantage and disadvantage. Fortunately there are multiple customizations of TEI, including TEI Lite for the elements sufficient for simple documents. TEI Lite is the most widely used TEI customization. TEI is popular in digital humanities, which also indicates it as a good option for preservation of texts. More information on TEI can be found in section 0.

PDF/A is an ISO standard dedicated for archiving various types of documents in digital form. The format is relatively new therefore it is not widely indicated as a master file format, neither by existing recommendations nor by current practices gathered by Succeed survey. Nevertheless it is based on PDF, which is very popular and also used for master file by 23% of survey respondents. Therefore it is reasonable at least for those who already use it to move from regular PDF to PDF/A. It is very important to

[49] ICC stands for International Color Consortium
[50] http://www.openjpeg.org/
[51] https://github.com/openplanets/jpylyzer
[52] http://iiif.io/

distinguish PDF/A from the PDF format. PDF/A is an archival format, which is based on PDF, but introduces specific restrictions/requirements to ensure appropriate visual representation of the document and other characteristics. For example it requires fonts to be embedded in the document, ICC-color based profiles and disallows encryption. There are three consecutive versions of the PDF/A format, each having several conformance levels. The conformance levels include[53]:

- Level B – ensures appropriate visual appearance of the document. This level has been introduced in the PDF/A-1 version.
- Level A – builds on level B, but in addition requires structured information about the document. This level has been introduced in the PDF/A-1 version.
- Level U – ensures that the text in the document can be extracted and appropriately interpreted. This level has been introduced in the PDF/A-2 version.

Also consecutive versions of the format added new capabilities to the format. The most important aspects of each version are:

- PDF/A-1 – introduces restrictions related to fonts, colors, etc.
- PDF/A-2 – introduces possibility to have different layers in the document, allows JPEG2000 compression and attachments to the document.
- PDF/A-3 – makes the attachments mechanism more flexible.

None of the versions is obsolete therefore all of them can be used for archiving purposes. They simply provide different set of features, which can be used, and different sets of conformance levels.

An alternative format for text representation is Unicode plain text file (encoded with UTF-8). The reason for it to be alternative is lack of support for structural information, as the file simply represents stream of characters. We recommend using UTF-8 encoding as it is compatible with ASCII and is able to encode various diacritics. It is also worthwhile to use normalized forms[54] of UTF-8 to store text files.

In case of historical documents, especially those with special characters not currently available in the Unicode standard, we recommend using MUFI specification (code points). Such an approach will minimize the risk of code point collisions between textual resources coming from different digitization projects or software tools. It is also likely for MUFI characters to be incorporated into the Unicode itself (e.g. 152 of MUFI characters were added to the Unicode 5.1). For details on MUFI please see section 0.

Descriptive metadata format

Recommended: DCMES (Dublin Core), MODS
Alternative: MARC21

[53] http://www.pdflib.com/knowledge-base/pdfa/
[54] http://en.wikipedia.org/wiki/Unicode_normalization

The most popular descriptive metadata format is Dublin Core (the full name is Dublin Core Metadata Element abbreviated as DCMES), which is globally recognized ISO standard. 71% of existing recommendations and 59% of survey respondents has indicated it as the main format for descriptive metadata in the context of long-term preservation. It is a simple and easy to use XML-based format. The simplicity of DCMES is an advantage and disadvantage at the same time. It is good because thanks to simplicity many institutions can easily use it. It is bad because the meaning of particular elements in the standard is not strict, which may cause various misunderstandings. If more detailed description is needed Dublin Core Metadata Initiative Terms (DCTerms) can be used, as those include all the elements from DCMES, and add additional ones, which allow for more precise description.

MODS format is quite popular with relatively high adaptation in the user community (16% of respondends use it for preservation, 47% of existing recommendations indicate it as a good option). MODS is based on XML, it can contain a richer description than Dublin Core, and is also based on MARC21 (though is not able to carry full MARC21 records), therefore can be easily created from existing MARC21 records.

MARC21 was also indicated in existing recommendations and survey. Nevertheless it is not highly recommended as it has several issues with interoperability. It has a specific encoding scheme for transportation purposes (MARC21 communication format), but it is not simple, it is not self-descriptive and definitely it is not human-readable. Additional complication is the possibility to encode MARC21 records using different encodings. It may cause additional issues, as for instance the offsets indicated in MARC21 leader (header) depend on characters and not bytes (and some characters can occupy more than one byte – depending on the encoding). It means that encoding needs to be know beforehand (before processing) and it is not available in the file itself. Because of these reasons the MARC21 format is proposed as alternative.

Structural metadata format

Recommended: METS

For structural metadata the only option is METS format. In practice there is no real alternative for the format. It is already used by 36% of survey respondents and it is indicated by existing recommendations in 59% of cases. It is an XML-based open standard, simple to apply and supporting various specific formats, including MODS, ALTO, TextMD, MIX and PREMIS (which are all recommended by Succeed project) . It is therefore the best option (and in practice the only one) to be used for structural metadata for long-term preservation.

Administrative metadata format

Recommended: PREMIS, MIX, TextMD

In case of administrative metadata existing recommendations and survey respondents indicate PREMIS for preservation and MIX or NISO Z39-87 for technical metadata of still images. TextMD is recommended as a technical metadata format for textual documents.

MIX is and XML-based format and the most popular implementation of the NISO Z39-87 standard. It can be also easily integrated with METS. It is therefore recommended for storing technical metadata about still images. PREMIS is in fact the only format used in practice to store preservation metadata. 41% of existing recommendations and 22% of survey respondents has indicated it. PREMIS can be also easily integrated with METS format, as it is XML-based. It is actively developed (currently the Editorial Board works towards version 3.0) and has its own PREMIS ontology for information exposure over semantic technologies. TextMD is not widely used by institutions from the survey. It is also not largely pointed by existing recommendations. If fact no indications are given for technical metadata of textual documents. This is why it seems to be a reasonable option to use a format, which is already well-integrated with structural metadata recommendations or preservation recommendations. TextMD is such a format – it is XML-based format and can be easily used in METS format as well as in PREMIS. It is also supported by characterization tools (e.g. JHOVE[55]).

OCR results format

Recommended: ALTO, PAGE
Alternative: UTF-8 encoded plain text

ALTO format has been indicated by 29% of existing recommendations. It is a format, which was developed to extend METS in order to provide both information about coordinates (ALTO format) as well as structural information (METS). The benefits and disadvantages of ALTO have been pointed in section 0. The main advantages include interoperability, readability (XML-based) and simplicity. The main disadvantages are related to limited number of supported region types and lack of support for capturing logical structure (this needs to be done by format container like METS). The ALTO format exports are also supported by some of the commercial OCR engines and is also a selection for ongoing initiatives (e.g. Europeana Newspapers project).

One of the main design goals of the PAGE format was to enable detailed and accurate description of any information which can be derived from a given document image, by overcoming limitations of existing formats (like ALTO) and allowing its use in applications requiring a very precise content representation (such as performance evaluation). The PAGE format does not have wide range of users, but it gains more and more attention, as it is used in such initiatives and projects like IMPACT Centre of Comptence[56], eMOP[57], Europeana Newspapers[58] or Transcriptorium[59].

[55] http://sourceforge.net/projects/jhove/
[56] http://www.digitisation.eu/

The alternative format is a simple text file encoded with UTF-8. The reason for it to be alternative is lack of support for structural information, as the file is simply stream of characters. We recommend using UTF-8 encoding as it is compatible with ASCII and is able to encode various diacritics. It is also worthwhile to use normalized forms of UTF-8 to store OCR results in such text files.

In case of historical documents, especially those with special characters not currently available in the Unicode standard, we recommend using MUFI specification (code points) to be used when training OCR engine (which results in MUFI characters in OCR output). Such an approach will minimize the risk of code point collisions between textual resources coming from different digitization projects or software tools. It is also likely for MUFI characters to be incorporated into the Unicode itself (e.g. 152 of MUFI characters were added to the Unicode 5.1). . For details on MUFI please see section 0.

5.2 Online delivery

Delivery file format

Recommended: JPEG, PDF, JPEG2000 (JP2), ePUB, MOBI derived from ePUB

Delivery files are for the end user - should be easy to use and simple to display them. It is also worthwhile to consider using several delivery formats for specific digital objects, as different users can have different preferences.

JPEG format has been indicated by most of existing recommendations (82%) and majority of Succeed survey respondents (71%). It is a general purpose image format which uses lossy compression to minimize the size of an image. JPEG is supported by practically all web browsers, including mobile ones.

PDF format is the most popular among Succeed survey respondents (77%) and it is also very popular in existing recommendations (53%). PDF is very popular, but requires special software tools to be displayed on the computer device. Some web browsers have lately added build-in support for PDF (e.g. Firefox and Chrome), so in some cases it is not a barrier anymore. PDF has also support for progressive download. It also supports multiple layers, therefore can be used for images or textual content or both.

JPEG2000 is also an option to consider for online delivery, especially when one wants to provide high-resolution images. JPEG2000 supports tiles and various resolution levels; therefore it is a perfect format for such application. It requires dedicated software tools to display in a user web browser, but there are already tools supporting such features

[57] http://idhmc.tamu.edu/emop/
[58] http://www.europeana-newspapers.eu/
[59] http://transcriptorium.eu/

(e.g. IIIF[60], OpenSeadragon[61]). Thanks to such solutions it is possible to use production master files as a direct source for online delivery of digital content.

For ebook readers textual format is required. The most popular formats in this context are ePub and MOBI, therefore those two formats are recommended in such cases. ePub and MOBI can be directly converted from ALTO, PAGE or UTF-8 encoded plain text. In case of MOBI format it is important to note that it is a proprietary format. The reason for recommending it is that it is the format supported by the Kindle® devices, which are currently very popular in the context of e-book readers. The best approach for using MOBI is to keep ePub as a primary delivery format and convert it (free tools are available) to MOBI to support wide range of users and their devices.

OCR results can be provided either together with the presentation format, e.g. in PDF or as a separate file, which is in format used for OCR results preservation.

Descriptive metadata format for online delivery

Recommended: DCMES (Dublin Core), EDM

Dublin Core Metadata Element Set (DCMES) is a must for each institution that wants to provide descriptive metadata online. It is a basic set of 15 elements, providing general information about the resource. Dublin Core is the most popular metadata format provided online by Succeed survey respondents (69%). It is the only format necessary to be supported when implementing OAI-PMH communication (OAI-PMH is widely accepted metadata harvesting protocol, used by Europeana and Digital Public Library of America). Although it is simple and very popular the main disadvantage is lack of precise interpretation of each element. This may cause inconsistency, e.g. on a level of metadata aggregator.

Europeana Data Model (EDM) has been introduced to enable delivery of richer information to Europeana portal than in case of Dublin Core or Europeana Semantic Elements. EDM was prepared to support all of important requirements from cultural heritage institutions. The idea was to increase interoperability of metadata, leverage semantic technologies, and provide finer granularity and more semantics. The EDM is based on existing formats and standards, such as Dublin Core, SKOS, and OAI-ORE. It is also already used by 22% of survey respondents. It is highly recommended for European institutions to use EDM for exposing metadata about provided content, as thanks to EDM the integration with Europeana is fully possible.

Identification of objects

Recommended: OAI Identifier, DOI

[60] http://iiif.io/
[61] http://openseadragon.github.io/

In the context of identifiers there are two main options: OAI Identifier and DOI.

The OAI Identifier is a free solution, which is based on domain names and provides possibility to implement persistent identifiers in repositories, which support OAI-PMH protocol. It does not build on a common infrastructure – it relies on the digital repository which implements OAI-PMH protocol and provides OAI Identifiers in OAI-PMH communication. It relies on domain names, which means that one part of the OAI Identifier contains domain name of the service providing OAI-PMH functionality. As a consequence it may introduce some confusion, e.g. when domain name is changed.

DOI is a paid service for keeping persistent identifiers of digital content. DOI is based on the Handle System and used by multiple of institutions (15% of respondents). DOI has been selected over the Handle System (which is also used by 15% of respondents) because it adds additional features, including persistence, consistency and robust technical infrastructure. The benefit of such an approach is reliable and existing infrastructure (provided by Handle System and DOI) as well as independence of specific technology (as opposed to OAI-PMH which is based on domain names).

5.3 Advanced and supporting technologies

Advanced and supporting technologies in the digitization related activities have a potential to improve interoperability, processing time and quality of the whole digitization workflow. We have investigated three aspects: semantic technologies, linguistic resources and tools packaging.

Linked Open Data

Recommended: RDFa, SPARQL

The Linked Open Data (LOD) paradigm introduces a new way of thinking about resources available on the web. The main idea behind LOD is to have the resources interlinked with other resources, so that it is easy to discover new resources and find relations between them. The term open suggests to have the data available using the open licenses, such as Creative Commons 0 Public Domain Dedication (which is used by Europeana). There are multiple standards related to semantic technologies, which can be used when publishing resources over the web. Those, which are maintained by the W3C[62] include RDF, OWL, SPARQL, RDFa, SKOS and RDFS.

We recommend investigating two standards when considering Linked Open Data: RDFa for representing RDF triples on the website and SPARQL for querying information available in RDF store. Both are maintained by the W3C. Obviously there are other

[62] http://www.w3.org/

options, which can be as well considered; nevertheless those two standards seem to be most appropriate for general purpose.

RDFa is a standard which makes it possible to embed RDF triples into HTML, XHTML or XML documents. RDFa enables an easy way for exposing resources and information in form of Linked Data. The features of RDFa can be used in a limited way (making implementation very simple – RDFa Lite) or fully, but then requiring more expertise (RDFa Core). As a result semantic information can be extracted from the website (e.g. from a digital library) by automated tools and then further processed. RDFa itself is already used by 21% of survey respondents, which is 62% of those who use semantic technologies.

In order to enable more advanced access to resources it is recommended to build a SPARQL interface for preserved data. SPARQL is a query language for RDF and a common way of accessing information stored in RDF (it is used by 8% of survey respondents, which is 23% of those who use semantic technologies). In order to provide SPARQL interface (endpoint) it is necessary to have an RDF datastore, which is a kind of database for RDF triples (also called triplestore). Such a datastore can be build up, for example, from information available on the web in RDFa standard.

Linguistic resources

Recommended: TEI, CMDI or LMF

For discovery, retrieval and reuse of linguistic data it is important that the data is stored in a predictable format. There are many elements that can be preserved in the context linguistic resources, we focus here on corpuses and dictionaries, which can be helpful when improving OCR techniques in the digitization workflow.

TEI format is primarily semantic rather than presentational; the semantics and interpretation of every tag and attribute are specified. Some 500 different textual components and concepts (word, sentence, character, glyph, person, etc.); each is grounded in one or more academic discipline and examples are given. TEI Lite is an XML-based file format for exchanging texts. It is a manageable selection from the extensive set of elements available in the full TEI Guidelines. TEI offers tools like ODD and ROMA, which assists a user in choosing a subset from the TEI repertoire. For linguistic resources special customization is available, called TEI Corpus. TEI is also already present in the cultural heritage sector. Therefore it is worthwhile to consider its use as well, especially for those who already use TEI for other purposes.

Component Metadata Infrastructure (CMDI) is developed within the CLARIN project. It provides a framework to describe and reuse metadata blueprints. Description building blocks ("components", which include field definitions) can be grouped into a ready-made description format (a "profile"). Both are stored and shared with other users in the Component Registry to promote reuse. Each metadata record is then expressed as an XML file, including a link to the profile on which it is based. The metadata is stored in

repositories which are harvested. CLARIN provides a central portal for discovery of resources (CLARIN Visual Language Observatory). Moreover, CLARIN makes special software available for editing CMDI records (Arbil). CLARIN aims to provide an infrastructure for research within Europe including libraries and public archives. This infrastructure will not be available to parties outside that domain like commercial enterprises and individuals.

Lexical Markup Framework is an ISO 24613:2008 standard. The goals of LMF are to provide a common model for the creation and use of lexical resources, to manage the exchange of data between and among these resources, and to enable the merging of large number of individual electronic resources to form extensive global electronic resources. Types of individual instantiations of LMF can include monolingual, bilingual or multilingual lexical resources. The same specifications are to be used for both small and large lexicons, for both simple and complex lexicons, for both written and spoken lexical representations. The linguistics constants like /feminine/ or /transitive/ are not defined within LMF but are recorded in the Data Category Registry (DCR) that is maintained as a global resource by ISO/TC37 in compliance with ISO/IEC 11179-3:2003. And these constants are used to adorn the high level structural elements. LMF is relatively new, but has already gained considerable popularity. According to some linguists the standard is not strict enough. ISO addressed that issue by creating reference structures for several subdomains.

Tools packaging

Recommended: package tools for targeted operating systems (at least for MS Windows and Linux)

Tools packaging is one of the elements that makes the maintenance and uptake of new tools easier. The benefit of having specific packages for certain operating systems is that the installation process can be automated. For example in case of Linux systems packaging provides means to install or update software packages, including shortcuts and command line tools. It can also automatically add documentation (e.g. to manpages). This would not be possible without a software package (although it would be possible to simply run a software from binaries, but then with no deep integration with the operating system). It is therefore highly recommended to use tools packaging techniques in order to deliver software to the end users. From the survey analysis it seems that MS Windows is the most popular operating system (87% of respondents). Linux is the second one (49%). Unix and MacOS have approx. 10% popularity each. This indicates that when building software packages at least MS Windows and Linux should be supported, so that most of the potential users can use automated installation procedure.

6. SUMMARY

This report provides a set of recommendations for specific aspects of digitization activities, such as master files creation, online delivery of digital objects or advanced and supporting technologies. This report provides also an overview of existing recommendations and current practices in the context of digitization activities, especially those related to text/printed documents. It also contains analysis of the survey conducted among digitization practitioners coming from various institutions across the globe. The recommendations were elaborated based on the overview of existing recommendations and survey analysis. It was done by identification of most suitable formats and standards for use in digitization related activities, with their advantages and drawbacks in mind.

Three summary tables of recommendations provide a concise view on selected options:

- Table 20 summarises recommendations for long-term preservation, including master file, metadata and OCR results formats.
- Table 21 summarises recommendations for online delivery, including delivery file formats, descriptive metadata formats and identification of objects.
- Table 22 summarises recommendation for advanced and supporting technologies, including Linked Open Data, linguistic resources and tools packaging.

Table 20 Recommendations for long-term preservation

Application	Recommended	Alternative
Master file format for still images	TIFF	JPEG2000 (JP2)
Master file format for textual documents	TEI, PDF/A	UTF-8 encoded plain text
Descriptive metadata format	DCMES, MODS	MARC21
Structural metadata format	METS	N/A
Administrative metadata format	PREMIS, MIX, TextMD	N/A
OCR results format	ALTO, PAGE	UTF-8 encoded plain text

Table 21 Recommendations for online delivery

Application	Recommended	Alternative
Delivery file format	JPEG, PDF, JPEG2000 (JP2), ePUB, MOBI derived from ePUB	N/A
Descriptive metadata format	DCMES, EDM	N/A
Identification of objects	OAI Identifier, DOI	N/A

Table 22 Recommendations for advanced and supporting technologies

Application	Recommended	Alternative
Linked Open Data	RDFa, SPARQL	N/A
Linguistic resources	TEI, CMDI, LMF	N/A
Tools packaging	At least MS Windows and Linux packages	N/A

There are also several interesting aspects that have been revealed by Succeed survey. Firstly, it is visible that there is still lack of quality assurance activities for OCR results, both for character and layout recognition (approx. 60-70% of respondents do not perform quality checks of OCR results, see Figure 28 and Figure 27). This issue is highly important to consider, as the textual resources are very valuable in the context of research (e.g. in humanities) and resource discovery (e.g. search engines). Secondly, it is also clearly visible that usage of advanced technologies (e.g. Named Entities Recognition, Geolocation) is very limited. More than 90% of respondents do not use such techniques at all (see Figure 25). It is clearly the aspect of digitization, which could be improved. Similarly, Linked Open Data are not leveraged by most of the respondents, as 66% of respondents do not use such technologies in their digitization activities (see Figure 23). It is also interesting that very small part of respondents are willing to provide master files online (29%). Finally, the survey showed that most of the respondents (55%) do not use digitization workflow management systems (see Figure 30). This means that the management of digitization processes and projects are handled manually in most of institutions. Professional digitization workflow management systems could provide automated and self-managed environment for digitization activities and therefore improve efficiency and quality of results.

BIBLIOGRAPHY

A. Antonacopoulos, C. C. (2013). ICDAR2013 Competition on Historical Newspaper Layout Analysis - HNLA2013. *Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR2013)* (pp. 1486-1490). Washington DC, USA: IEEE.

S. Pletschacher, A. A. (2010). The PAGE (Page Analysis and Ground-Truth Elements) Format Framework. *Proceedings of the 20th International Conference on Pattern Recognition (ICPR2010)* (pp. 257-260). Istanbul, Turkey: IEEE-CS Press.

GLOSSARY OF ACRONYMS

ALTO	Analyzed Layout and Text Objects – standard for technical metadata for Optical Character Recognition (OCR). See: http://www.loc.gov/standards/alto/
ANSI/NISO Z39.87	Standard which defines a set of metadata elements for raster digital images. See: http://www.niso.org/kst/reports/standards?step=2&gid=None&project_key=b897b0cf3e2ee526252d9f830207b3cc9f3b6c2c)
ASCII	American Standard Code for Information Interchange (ANSI X3.4-1986).
CIDOC CRM	CIDOC Conceptual Reference Model for describing concepts and relationships in cultural heritage documentation See: http://www.cidoc-crm.org/
CMDI	Component MetaData Infrastructure. See: http://www.clarin.eu/cmdi
copyrightMD	Format for representing copyright metadata. See: http://www.cdlib.org/groups/rmg/
DCMES	Dublin Core Metadata Element Set – standard for resource description. See: http://dublincore.org/documents/dces/
DCTerms	Dublin Core Metadata Initiative Metadata Terms See: http://dublincore.org/documents/dcmi-terms/
DjVu	File format especially prepared for scanned documents. See: http://www.caminova.com/docs/techinfo/DjVu3Spec.pdf
DNG	Digital Negative image coding format for storing camera raw files. See: http://wwwimages.adobe.com/www.adobe.com/content/dam/Adobe/en/ products/photoshop/pdfs/dng_spec_1.4.0.0.pdf)
DOI	Digital Object Identifier provides services for registering of persistent interoperable identifiers. See: http://www.doi.org/
Dublin Core	see DCMES
EAD	Encoded Archival Description – standard for the encoding of finding aids. See: http://www.loc.gov/ead/
EDM	Europeana Data Model See: http://pro.europeana.eu/edm-documentation
ePUB	Electronic Publishing format. See: http://idpf.org/epub/30
GIF	Graphics Interchange Format image coding standard. See: http://www.w3.org/Graphics/GIF/spec-gif89a.txt
JP2	See JPEG2000
JPEG	Image coding standard. When mentioned in the document it usually references to the JFIF standard. See: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detai

	l.htm?csnumber=54989)
JPEG 2000	Image coding standard using wavelet-based compression method. Whenever mentioned in the document it references to JP2 format – JPEG 2000 Part 1 standard – core coding system. See: http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=37674)
LIDO	Lightweight Information Describing Objects – format intended for delivering and connecting data on the web. See: http://network.icom.museum/cidoc/working-groups/data-harvesting-and-interchange/what-is-lido/).
LMF	Lexical Markup Framework. See: http://www.lexicalmarkupframework.org/
MARC	A set of formats for bibliographic and related information. See: http://www.loc.gov/marc/
MARC21	Format for bibliographic data representation. See: http://www.loc.gov/marc/bibliographic/
METS	Metadata Encoding & Transmission Standard – standard for encoding descriptive, administrative and structural metadata See: http://www.loc.gov/standards/mets/
MIX	NISO metadata for images in XML Schema – set of technical metadata elements for images, which provides format for interchange/storage of the data specified in ANSI/NISO Z39.87. See: http://www.loc.gov/standards/mix/
MOBI	Mobipocket proprietary format. See: http://www.mobipocket.com/dev/
MODS	Metadata Object Description Schema – schema for bibliographic element set. See: http://www.loc.gov/standards/mods/
MPEG-21 DIDL	Representation of the Multimedia framework (MPEG-21) Part 2: Digital Item Declaration See: http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=41112)
MrSID	Multi-resolution seamless image database – format for encoding georeferenced raster graphics. Proprietary format owned by LizardTech. See: http://www.lizardtech.com/
MUFI	Medieval Unicode Font Initiative. See: http://www.mufi.info/
OAI Identifier	Identifier format to provide persistent identifiers in repositories that implement OAI-PMH. See: http://www.openarchives.org/OAI/2.0/guidelines-oai-identifier.htm
OAI-ORE	Open Archives Initiative Object Reuse and Exchange – standard for description and exchange of aggregations of Web resources. See: http://www.openarchives.org/ore/
ObjectID	International standards for describing art, antiques and antiquities .

	See: http://archives.icom.museum/objectid/
OCR	Optical Character Recognition. See: http://en.wikipedia.org/wiki/Optical_character_recognition
PAGE	Page Analysis and Ground-Truth Elements. See: http://www.primaresearch.org/tools.php
PDF	Portable Document Format – format for coding electronic documents. See: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=51502
PDF/A	Version of Portable Document Format specialized for long-term preservation. See: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=38920 See: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=50655 See: http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?ics1=37&ics2=100&ics3=99&csnumber=57229)
PNG	Portable Network Graphics image coding standard. See: http://www.w3.org/TR/2003/REC-PNG-20031110/)
PREMIS	Preservation Metadata Maintenance Activity – data dictionary and format for preservation metadata. See: http://www.loc.gov/standards/premis/
PSD	Photoshop Document proprietary format for coding Adobe Photoshop® documents. See: http://www.adobe.com/devnet-apps/photoshop/fileformatashtml/
RAW	A camera raw image format. See: http://en.wikipedia.org/wiki/Raw_image_format
RDFa	Resource Description Framework in Attributes. See: http://www.w3.org/TR/xhtml-rdfa-primer/
SPARQL	SPARQL protocol and RDF query language. See: http://www.w3.org/TR/rdf-sparql-query/
TEI	Text Encoding Initiative which develops standards for representation of texts in digital form. In the documents it usually references to TEI guidelines. See: http://www.tei-c.org/Guidelines/
textMD	Technical Metadata for Text – format for specifying technical metadata for text-based digital objects. See: http://www.loc.gov/standards/textMD/
TIFF	Tagged Image File Format image coding standard. See: http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf)
Unicode	Standard for encoding, representation and handling of text. See: http://www.unicode.org/
UTF-8	Universal Character Set Transformation Format – 8-bit – character encoding for Unicode, compatible with ASCII See: http://www.unicode.org/versions/Unicode6.0.0/ch03.pdf
XML	Extensible Markup Language – flexible text format for electronic

	information exchange. See: http://www.w3.org/XML/
XMP	Extensible Metadata Platform – format for metadata which can be embedded into described file. See: http://www.adobe.com/products/xmp/

ATTACHEMENT A. SUCCEED SURVEY QUESTIONNAIRE

Below you will find Succeed questionnaire. It is in form of screenshots for presenting the same look and feel that respondents had.

Succeed survey: digitisation standards&formats

* Required

Basic information

Questions related to basic information about the institution filling in the questionnaire.

1. Please specify the name of the institution (and optionally division) you represent. *

If there are multiple divisions dealing with digitisation at your institution and you represent only one of them, please specify your division name.

This is a required question

2. What is the type of your institution? *

Please specify the type of your institution in the context of holdings. In case of multi-type institution, mark all relevant options.

- [ ] Gallery
- [ ] Library
- [ ] Archive
- [ ] Museum
- [ ] Other:

3. What is the number of employees working with digitisation in your institution? *

Please specify the number of employees in your institution involved in digitisation activities. Please select an appropriate option or provide an exact number in 'Other' option.

- < 5
- 5 - 10
- 11 - 50
- 51 - 100
- > 100
- Other:

4. Please provide your e-mail address for any follow up questions.

We would like to obtain your e-mail address to have the opportunity to ask you follow up questions if some issues will not be clear for us.

5. Would you like to receive results of our analysis related to this survey? *

We would like to send you the results of our analysis related to this survey, if you are interested, please mark yes. If you do not want to be bothered, please mark no. The 'Other' option is for conditional interest or different e-mail address to send the results to.

- Yes
- No
- Other:

« Back Continue »

* Required

Long term preservation

Questions related to formats and standards used in the context of long term preservation.

6. What are the master file formats you use for preservation of images? *

Please select all the formats you use for master files preservation. If you store master files in several formats, please check them all (or add in 'Other'). If part of the files are stored in one format and another part in another one, then please check them all (or add to 'Other' option).

- ☐ TIFF
- ☐ JPEG 2000
- ☐ JPEG
- ☐ PDF
- ☐ Other:

7. Please provide detailed characteristics of the master file formats you use.

Please consider to provide the following characteristics for each format: format version, uncompressed or compressed data, lossy or lossless compression, resolution (e.g. in DPI), bit-depth for each color channel, color mode (e.g. RGB, CMYK), color profile (e.g. sRGB).

8. What are the metadata formats and standards you use for long term preservation of your digital objects? *

Please mark all the metadata formats that you use for long term preservation of your digital objects. If you use several formats for one object, mark them all. If you use different formats for different objects, please mark them all. In the brackets you will find the type of metadata. In case you use other formats or standards please list them all with their types in the 'Other' option.

- [ ] Dublin Core (descriptive)
- [ ] EAD (descriptive)
- [ ] LIDO (descriptive)
- [ ] MAB (descriptive)
- [ ] MARC (descriptive)
- [ ] METS (packaging)
- [ ] MODS (descriptive)
- [ ] XMP (descriptive)
- [ ] MPEG-21 DIDL (packaging)
- [ ] OAI-ORE (packaging)
- [ ] PREMIS (preservation)
- [ ] NISO Z39.87 (technical for still images)
- [ ] textMD (technical for text)
- [ ] Other:

9. What are the file formats you use for storing text representation (e.g. OCR results)? *

When digitising printed texts it is common to perform Optical Character Recognition (OCR) on them. If you preserve/store these OCR results, please mark the formats you use for this purpose. Use 'Other' to describe not listed options (e.g. if you use proprietary/non-standard XML schema).

- [ ] XML with coordinates information (e.g. ALTO)
- [ ] XML with structural information (e.g. TEI)
- [ ] Plain text
- [ ] PDF
- [ ] Other:

10. How many pages in digital form have you already preserved? *

Please note that a digital object can be composed of multiple pages. We ask for pages estimation to have better understanding of your experience in the field. If you do not know the number of pages, please specify statistics in the 'Other' option (e.g. number of digital objects).

- <10 000 pages
- 10 000 - 100 000 pages
- 100 000 - 1 000 000 pages
- > 1 000 000 pages
- Other:

11. How many pages do you approximately digitise/process each year (last 5 years)? *

Please let us know the number of pages you process each year using your digitisation workflow. We would like to know your usual workload in digitisation activities. Please take into consideration the last 5 years.

- < 1000
- 1000 - 10 000
- 10 001 - 100 000
- 100 001 - 250 000
- > 250 000
- Other:

12. Would you be willing to make the master files available online (part or all of them)? *

Sometimes master file copies are required for specific research tasks (e.g. OCR improvements), reproduction or other purpose (e.g. new format of surrogates/presentation version). We ask this question to understand your current policy related to master files online availability (under the same conditions as surrogates/presentation version). If the answer is not clear in your case, please use the 'Other' option and describe shortly your considerations.

- Yes
- No
- Other:

13. Which of the national or international recommendations/guidelines do you follow in the context of digitisation and digital preservation? *

Please specify the guidelines or recommendations that you follow. If you do not follow any please state why.

* Required

Online delivery of digital objects

Questions related to formats and standards used in delivering digital objects to online users (e.g. via digital library portal).

*14. What are the file formats of surrogates (presentation version) you use for online delivery? **

If you have a digital library portal (or similar website) you need to provide digital objects in specific formats. Please mark all formats you use for delivering digital objects to internet users.

- [ ] DjVu
- [ ] ePub
- [ ] GIF
- [ ] JPEG
- [ ] JPEG2000
- [ ] MOBI
- [ ] MrSID
- [ ] PDF
- [ ] PNG
- [ ] Other:

*15. What are the formats of metadata you provide to the internet users or aggregation services? **

If you provide metadata online, either for internet users on a website or aggregation services using dedicated protocols such as OAI-PMH, please mark all the formats you make publicly available.

- [ ] Dublin Core Metadata Element Set (set of 15 basic elements)
- [ ] DCMI Metadata Terms (extended set of descriptive metadata)
- [ ] Europeana Data Model
- [ ] Europeana Semantic Elements
- [ ] CIDOC CRM
- [ ] FRBRoo
- [ ] JSON-LD
- [ ] I do not provide metadata over the internet
- [ ] Other:

16. What functionality do you provide to your users for searching full text (e.g. OCR results) online? *

If you provide text representation (e.g. OCR results) for the end users or external systems, then please mark relevant forms.

- [ ] Full Text search with keyword highlighting on image
- [ ] Full Text search with OCR text displayed
- [ ] Full Text search embedded into viewer (e.g. PDF)
- [ ] Full Text available as download
- [ ] I do not provide OCR results online
- [ ] Other:

17. How many pages in digital form have you already published online? *

Please note that a digital object can be composed of multiple pages. We ask for pages estimation to have better understanding of your experience in the field. If you do not know the number of pages, please specify statistics in the 'Other' option (e.g. number of digital objects).

- ( ) <10 000 pages
- ( ) 10 000 - 100 000 pages
- ( ) 100 000 - 1000 000 pages
- ( ) > 1 000 000 pages
- ( ) Other:

18. Which of the national or international recommendations/guidelines do you follow in the context of online availability of digital objects? *

Please specify the guidelines or recommendations that you follow in the context of online availability of digital objects in your institution. If you do not follow any please state why.

* Required

Emerging standards, formats and approaches

Questions related to new and innovative approaches that become important in the context of digital imaging and cultural heritage.

19. Which formats/standards/protocols do you use to provide your digital objects using Linked Data approach? *

It is possible to provide your data in such a way that particular objects are linked to others using such technologies as RDF*, HTTP, URI. If you provide your digital objects as Linked Data, please specify formats/standards/protocols you use in this context.

- [ ] RDFa
- [ ] SPARQL
- [ ] POWDER
- [ ] I do not provide content using Linked Data
- [ ] Other:

20. Which of the persistent identifiers do you use for your digital objects?

Each digital object should have a persistent identifier, so that it can be identified and discovered in the future. Please mark all the relevant options you use for objects identification.

- [ ] DOI
- [ ] Handle System
- [ ] OAI Identifier
- [ ] ORCID

21. Which advanced technologies or tools do you use to enhance/enrich your digital content?

It is possible to enrich your digital content with advanced technologies, e.g. using Named Entities Recognition or geolocation. If you use any of such technologies or tools, please describe it shortly in the textbox below.

22. Which formats or standards do you use to store linguistic resources for your OCR enigne? *

It is possible to improve OCR results using word dictionaries or lexicons. Please specify the formats or standards that you use to store such information (e.g. dictionaries).

- [ ] CMDI
- [ ] LMF
- [ ] TEI
- [ ] I do not store linguistic resources
- [ ] Other:

23. What is your approach in the context of OCR evaluation for text recognition? *

After OCR is done on your digital objects it is possible to verify how good the OCR tool performs in the context of text recognition. It can be related for example to precision of OCR on character level.

- [ ] No checking
- [ ] Based on OCR engine statistics (e.g. confidence level)
- [ ] Dedicated evaluation activities (own method)
- [ ] Other:

24. What is your approach in the context of OCR evaluation for layout recognition?

After OCR is done on your digital objects it is possible to verify how good the OCR tool performs in the context of layout recognition. It can be related for example to precision of OCR image/table identification.

- [ ] No checking
- [ ] Based on OCR engine statistics (e.g. confidence level)
- [ ] Dedicated evaluation activities (own method)
- [ ] Other:

25. Which emerging/innovative technologies, standards, formats do you use or plan to use in the context of digital objects? *

If you use or plan to use any innovative or new standard, format or technology for any purpose related to your digital objects, please describe it in the textbox below.

« Back Continue »

* Required

Standards in digitisation related tools

Questions related to standards in the context of tools/software packages you use in the digitisation workflow. You might need assistance of your IT staff to answer these questions.

*26. Which are the Operating System families used in your institution in digitisation workflow? **

We would like to know what is the technical environment (operating systems) of your digitisation workflow, e.g. OS used in servers with digital library or preservation system. It is important in terms of deployment of new digitisation tools.

- ☐ Linux
- ☐ MacOS
- ☐ MS Windows
- ☐ Unix
- ☐ Other:

*27. Do you have any preference related to digitisation tools packaging or web services availability? **

When adding a new tool or web service to your digitisation workflow it is always a matter of deployment and integration with existing environment. If you have any preferences in this context please specify them in the texbox below. For example, let us know if you prefer to have .exe / .msi / .deb / installation packages, or maybe you prefer to have web services exposed as REST. Maybe you have completely other requirements related to new tools integration?

28. Which digitisation workflow management systems do you use?

Please specify the name of the digitisation workflow management system used by your institution to perform and control digitisation activities.

- [ ] Goobi
- [ ] docWorks
- [ ] dLab
- [ ] I do not use digitisation workflow management system
- [ ] Other:

29. Which digital library systems do you use for online presentation of your digital objects?

Please mark all relevant systems (or use 'Other' option if not listed) you use to make your digital objects available online.

- [ ] DSpace
- [ ] EPrints
- [ ] Greenstone
- [ ] Other:

30. Do you provide any kind of application programming interface (API) to your data/digital content?

If you provide an API to your digital content (e.g. as Europeana does via Europeana API), please describe how you do that, what standards you follow, etc.

« Back Submit