

# DELIVERABLE

**Project Acronym:** Europeana Newspapers  
**Grant Agreement number:** 297380  
**Project Title:** A Gateway to European Newspapers Online

---

## **D2.2 Specification of requirements of OCR and structural refinement-services for digitised newspapers in Europeana**

---

**Revision:** 1.0  
**Authors:** Clemens Neudecker, KB

**Contributions:** WP2 participants

<b>Project co-funded by the European Commission within the ICT Policy Support Programme</b>		
<b>Dissemination Level</b>		
P	<b>Public</b>	x
C	<b>Confidential, only for members of the consortium and the Commission Services</b>	

## Revision History

Revision	Date	Author	Organisation	Description
0.1	08-11-2012	Clemens Neudecker	KB	Created
0.2	14-01-2013	Clemens Neudecker	KB	Updated
0.3	06-02-2013	Clemens Neudecker, Lotte Wilms	KB	Comments NLF, BnF, CCS addressed
0.4	21-02-2013	Knut Lohse	SBB	Internal review
1.0	21-02-2013	Clemens Neudecker	KB	Final version after internal review

### Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

## Table of Contents

1. Executive Summary .....	4
2. Technical requirements.....	5
2.1 General requirements.....	5
2.2 Requirements for OCR.....	6
2.3 Requirements for OLR .....	8
2.4 Requirements for NER .....	8
3. Delivery requirements .....	11
3.1 Content selection.....	11
3.1.1 Master List .....	11
3.2 Data preparation.....	15
3.3 Delivery packaging .....	16
3.4 Tools.....	17
3.5 Delivery .....	18
4. Conclusion .....	20

## 1. Executive Summary

The main objective of Work Package 2 is to enhance and refine already digitised newspaper pages as part of the general aggregation process in the Europeana Newspapers Project. Apart from full-text for about 10 million pages, also millions of single articles with related metadata and named entities (persons, geo names, etc.) will be automatically detected, tagged and packaged for delivery to Europeana. In this way the user experience for searching and retrieving newspapers via Europeana will be dramatically enhanced compared to the current solutions.

This document outlines the technical and administrative requirements for applying structural refinement services to digitised newspaper collections in Europeana Newspapers, with a view on further integration into Europeana services.

At first, the technical requirements are specified for the three distinct refinement processes in the project, namely Optical Character Recognition (OCR), Optical Layout Recognition (OLR) and Named Entities Recognition (NER).

The following section explains how the delivery of content from the libraries to the refinement partners is organised in order to guarantee a smooth and sound processing at scale.

Separate sub-sections on the file naming and directory structure requirements follow, as well as an explanation of the various software tools and technologies that are used to support the process.

## 2. Technical requirements

### 2.1 General requirements

Structural refinement of digitised documents (and newspapers in particular) is a complex and often challenging task that can be broken down into various sub-tasks.

Typically the whole process starts out with image capture, with the next step being image enhancement, which again includes various sub-processes. Cropping, straightening of text lines and binarisation (the transformation into a black-and-white image) are some exemplary pre-processing steps before segmentation is applied. Segmentation (also sometimes referred to as “zoning”) aims to hierarchically break down a document into distinct sections consisting of text elements and non-textual elements such as illustrations or control characters, then paragraphs, followed by lines, words and finally glyphs. Only then the actual text recognition process (OCR) is triggered, sometimes iteratively.

After the text has been detected and exported, (optional) processing steps include the verification of the recognised words against a dictionary and the further enrichment of the recognised text with a semantic layer, such as for example the tagging of standardised person or place names being mentioned in the text.

Given the complexity of the full pipeline, it is apparent that in itself it is rather sensitive to the nature of the input material. For example compressed or low-resolution images yield significantly worse results from what is usually referred to as “master images”, i.e. uncompressed high resolution images, so only master images will be accepted for the refinement workflows.

When speaking of a high resolution master image, optical resolution as in the spatial resolution of a digital image relating to the physical size of it is meant to be at least 300 ppi. This measure is typically indicated as DPI (“dots-per-inch”), or, more often, PPI (“pixel-per-inch”)<sup>1</sup>. Experience shows that the optimal resolution for refinement with OCR depends on several factors, such as the print size (small print size requires higher optical resolution), but also the number of colours in a scanned image, but should not be below 300 ppi to arrive at decent OCR results.

Generally there is a preference to use greyscale or even coloured images for the OCR – while the images are converted to bitonal (black-and-white) images internally as part of the text recognition process, this conversion is highly optimised for OCR purposes and does yield significantly better results that derived from a bitonal image to start with.

Also, the FP7 project IMPACT delivered a [“Best practice guide”](#) for image capture with a view on OCR - wherever possible, the selection of material from Europeana Newspapers content providing partners has been guided by the recommendations in that document.

While there is no strict limitation with regard to the image file formats that can be processed by the technical partners in the project, there is a strong bias towards proven standard image file formats in the cultural heritage domain that are also well supported by the various software applications used in the project. These include:

- TIFF (uncompressed, Group4, LZW)

---

<sup>1</sup> For a discussion of DPI vs. PPI, see:

[http://en.wikipedia.org/wiki/Dots\\_per\\_inch#DPI\\_or\\_PPI\\_in\\_digital\\_image\\_files](http://en.wikipedia.org/wiki/Dots_per_inch#DPI_or_PPI_in_digital_image_files)

- JPEG
- PNG
- JPEG2000 (.JP2)

## **2.2 Requirements for OCR**

Within the Europeana Newspapers Project, the University of Innsbruck (UIBK) is the main provider for Optical Character Recognition (OCR). Around eight million newspaper pages are foreseen for OCR through the University of Innsbruck as part of the project.

The work will be carried out within the “[Abteilung für Digitalisierung und elektronische Archivierung](#)” (Department for Digitisation and Digital Preservation”) who have extensive experience with OCR from participation in various European project such as [MetaE](#), [IMPACT](#) and others and are also currently providing the technical and administrative backbone of the European eBooks on demand service [EOD](#), which also includes OCR amongst its services.

The University of Innsbruck is making use of the state-of-the-art commercial application for OCR, [ABBYY's FineReader](#), which was also developed further as part of the IMPACT project. In the course of 2012, the University of Innsbruck have been modifying their OCR service platform to use the FineReader Engine SDK instead of the Recognition Server because it gives more flexibility in the configuration while maintaining aptitude for large-scale processing. The version of the FineReader SDK which is being used is 11.0, the most recent release to date.

Accordingly, the requirements for OCR are (to some extent) also defined by what the software that is being used supports.

The workflow for OCR processing at UIBK is supported by various software tools and explained in more detail in figure 1 below.

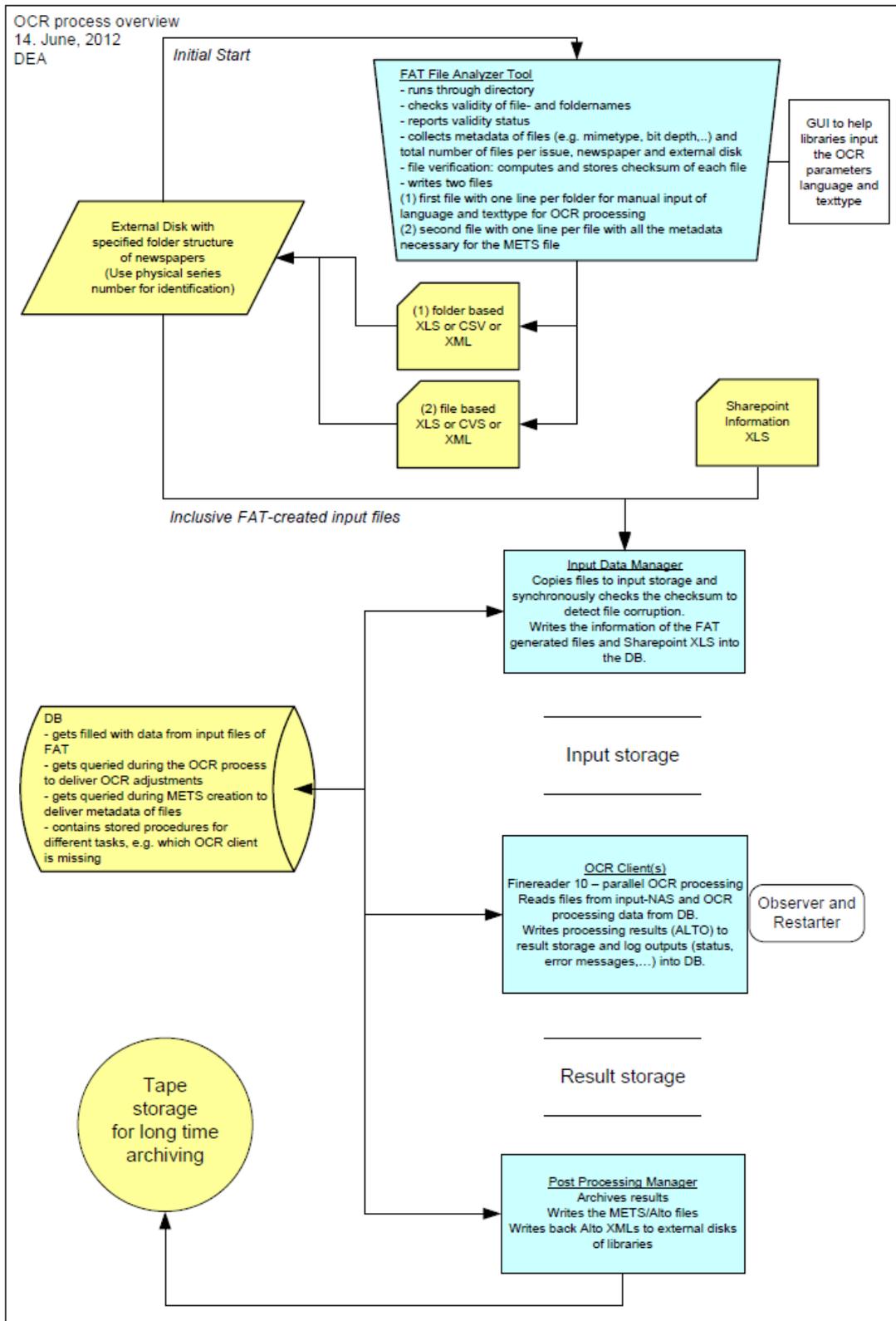


Figure 1: OCR processing at the University of Innsbruck

## 2.3 Requirements for OLR

In addition to the roughly eight million pages of “regular” OCR provided to the project by the University of Innsbruck, another two million pages will be OCR’ed with additional structural refinement (OLR = Optical Layout Recognition), such as the separation of articles and also page classification, by [Content Conversion Specialists](#) (CCS), Hamburg. CCS makes use of their [docWorks](#) software technology. This is an intelligent application for automatic conversion, structuring, and indexing of printed or electronic documents such as books, journals, newspapers and magazines. It can even locate and categorise key data from a variety of documents. Each automatic detection step can be followed by manual verification, depending on the quality level required. For mass digitisation projects, manual verification is typically reduced to a minimum, while boutique digitisation of small collections often allows for more semantic tagging and manual correction.

After raw data verification and ingest, the conversion process starts off with page analysis to determine page frames and correct alignment/angle of the page, followed by several zoning and structure recognition steps, where each element is assigned a specific zone category, and the individual elements (e.g. headlines, text blocks, illustrations) are grouped together into articles.

For the Europeana Newspapers Project, CCS proposed three different data processing scenarios to the partner libraries:

1. Conversion on site at the library (This requires the installation of the complete program, and would hence lead to the most pronounced changes to the library's environment. The advantage is that the partner library can gain invaluable experience in the full functionalities of docWorks and evaluate its benefit, to help decide whether to use it in future projects.)
2. Conversion offshore, final QA at the library (This solution means a smaller installation is needed, and the data is shipped to the library by internet transfer, for the final QA step only)
3. Conversion offshore, final QA at the library by backup shipment (Similar to scenario 2, the only difference being the mode of transport for the data. If the library has stringent security protocols that do not allow FTP file transfer, the data can also be shipped on hard drives)

Four out of the five partner libraries that CCS is working with in the OLR section (National Library of Estonia, Bibliothèque nationale de France, Staatsbibliothek Berlin, Staats- und Universitätsbibliothek Hamburg) chose scenario 2, the offshore conversion solution, with only the National Library of Finland opting for scenario 1, as they have been working with docWorks for many years already and have the complete software and hardware setup at hand.

## 2.4 Requirements for NER

For a subset of the OCR’ed content from partners in Dutch, English or German language, the National Library of the Netherlands will provide tools and technologies for the extraction of named entities such as person and place names or organisation. The NER system that is going to be implemented will build on preliminary work that was carried out in the IMPACT project. The [Stanford](#) NER tagger with the [adaptations from INL](#) is used and further enhanced for the project.

In the course of the first year, following the analysis of the content from libraries, it became obvious that only little content is going to be provided in English language (since the British Library is a networking partner only and does not provide digitised newspaper pages to the projects data set). While there is a possibility that the addition of further associated partners to the consortium in 2013 might add some English content, it has also been investigated in how far it would be feasible to

support a third language other than English by the NER workflow. This entails that e.g. training material (data already annotated with named entities), gazetteers (lists of normalised person or place names) or even a trained model for the NER system are already in place. From the investigation it followed that the only NER resource that could be directly taken up in the project without major extra effort is that provided for Latvian. Thus the processing of Latvian content with NER will therefore be further investigated in 2013.

However, given that a significant share (approx. 25%) of the full text produced by the project will be in French language, particular attention was given to investigating the feasibility of extending the NER software used in the project with support for French language, or even use another software tool designed for French NER. However, it turned out that the system having been developed by the Quaero<sup>2</sup> project is not yet advanced enough to deal with:

- a) Noisy OCR content (which may have word accuracy significantly lower than 90%)
- b) Historical language variation

In addition, there are no resources known to be currently available at either KB or BnF that could be used as a starting point for adapting the Stanford NER software for French.

In December 2012, the BnF has submitted a proposal to the Project Management Board for the development of NER resources for French via collaboration with the ACASA LIP6 Group of the Université Pierre et Marie Curie in Paris. The decision to take on NER for French is currently being investigated by the members of the Project Management Board.

During the first year, several NER software systems were being looked at to determine what would be the best fitting solution for the project. A simple evaluation was conducted between the following frameworks, looking at criteria such as languages supported, ease of use, throughput and suitability for processing OCR'd content:

- [Stanford \(standard\)](#)
- [Stanford \(KB version\)](#)
- [Stanford \(INL version\)](#)
- [OpenNLP](#)
- [AlchemyAPI](#)

Due to time constraints, the following systems had to be excluded from the analysis after a first high-level evaluation:

- [GATE](#)
- [Lingpipe](#)
- [NLTK](#)
- [Illinois NER](#)
- [Pendulum](#)

From the [evaluation](#) it followed that the best results can indeed be expected from the Stanford tagger. Also, the Stanford software has already been used extensively by the KB in former

---

<sup>2</sup> [Galibert et. al.: Extended Named Entity Annotation on OCR'd Documents:](#)

[From Corpus Constitution to Evaluation Campaign, LREC2012.](#)

projects, and enhanced with a spelling variation module (to deal with historical spelling) and a module for processing noisy OCR as input during the IMPACT project by the Dutch Institute of Lexicology in Leiden (INL), who are also further developing these components as part of the Dutch [NAMESCAPE](#) project. A decision was therefore taken to use the modified version of the Stanford system by INL and further adapt it to the purpose of the Europeana Newspapers Project at KB. One important modification is for example to support additional input formats.

An important issue that was raised in the first year of the project was the question of how the information derived from the processing with NER should be represented in the final output of the refinement process. The problem is that the [ALTO standard](#), which is employed as the main output format from the OCR/OLR processing in the project, does not currently provide the necessary features required for encoding of semantic information. On the other hand, almost all NER software provide their output in either [BIO-format](#) or just plain text with tags, both of which are unsuitable for the Europeana Newspaper Project because of the loss of coordinates information. The information about the coordinates of the words is essential for the later presentation in the newspaper browser developed by TEL.

Accordingly, a discussion was formed and an options paper provided by KB which outlines several alternative ways for storing the NER output.

The options include:

- BIO format
- TEI
- ALTO
- Separate DB/file
- Open Annotation

Other possibilities, such as representing the entities directly in the METS container have also been considered, but for now omitted from the options paper because of their inherent complexity.

Further discussion between technical partners and libraries will determine the final output format for the NER workflow, which will in turn be implemented by mid-2013.

A first prototype of the Named Entities Annotator tool has been made available on github in February 2013: <https://github.com/KBNLresearch/europeanp-ner>.

Finally, in the course of 2013 the granularity of information that can actually be derived from the NER processing chain will be determined precisely. For example, disambiguation of entities and referencing of authority files are more advanced ways to refine NER results, which in turn require a more sophisticated implementation (and require different presentation options as well).

## 3. Delivery requirements

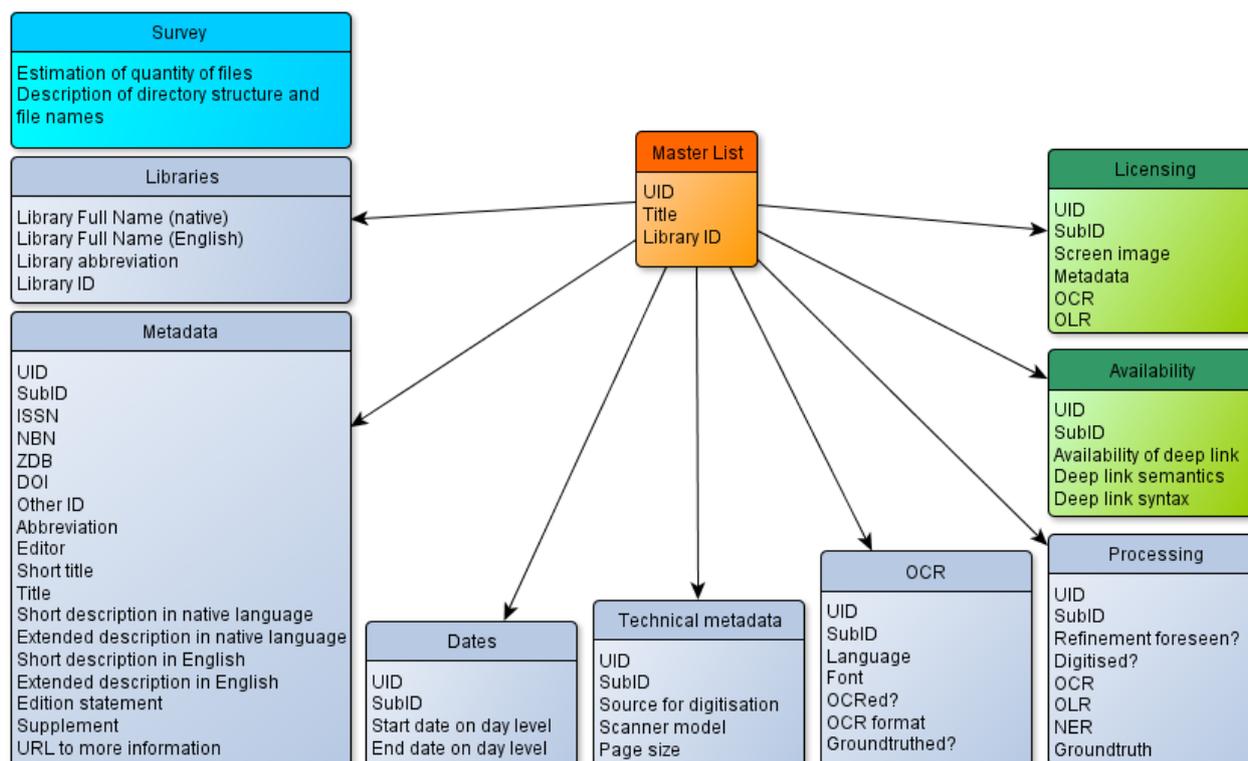
### 3.1 Content selection

Altogether more than 10 million files from 12 different libraries will be refined in Work Package 2. This demands that there are clear and transparent mechanisms in place for carrying out the content selection and for tracking progress.

#### 3.1.1 Master List

For this purpose, a central tracking list has been created which collects all the information required for the planning and tracking of the refinement procedures throughout the project. This so called “master list” holds the title level information about the newspapers selected by the libraries and is centrally managed on the project Sharepoint.

Initially, the technical partners of the project discussed the required fields for the list and thought of a database like structure, with one master list and several sub lists, each linking to the master list by means of a unique ID on the title level.



**Figure 2: Original tracking list design**

However, after discussion with the libraries in the [WP2 Workshop in Hamburg](#), several of the fields were decided to be irrelevant or redundant and could therefore be taken out of the overview. Due to this, the amount of information needed was much smaller than originally thought, thus making it possible to use only a single list instead of many separate ones. After several iterations of the list, the final columns are:

Column	Type	Description	Required
UID	Single line of text	Unique identifier of title. E.g. BnF_00001.	✓
Partner	Lookup	Library short name. E.g. BnF	✓
NewspaperTitle	Single line of text	Original newspaper title	✓
StartDate	Single line of text	Start year of title for ENP	✓
EndDate	Single line of text	End year of title for ENP	✓
Language	Choice	Tick box(es): <ul style="list-style-type: none"> <li>• Dutch</li> <li>• English</li> <li>• Estonian</li> <li>• Finnish</li> <li>• French</li> <li>• German</li> <li>• Latvian</li> <li>• LatvianGothic</li> <li>• Polish</li> <li>• SerbianCyrillic</li> <li>• SerbianLatin</li> <li>• Turkish</li> <li>• Other</li> </ul>	✓
FontType	Choice	Tick box(es): <ul style="list-style-type: none"> <li>• Normal</li> <li>• Gothic</li> <li>• Ottoman</li> </ul>	✓
PageSize	Choice	Choice: <ul style="list-style-type: none"> <li>- small</li> <li>- medium</li> <li>- large</li> </ul> Small = <10.000 characters per page Medium = 10.000 - 20.000 characters per page Large = >20.000 characters per	✓

		page	
Pages	Number	Total number of pages	✓
Source	Choice	Choice: - Original - Microfilm	✓
Refinement	Choice	Choice: - No refinement - OCR (UIBK) - OLR (CCS) - NER (KB)	✓
Digitised	Yes/No	Whether the title is already digitised.	
OCRed	Yes/No	Whether the title is already OCRed.	
MasterFormat	Choice	Tick box(es): • TIFF colour • TIFF greyscale • TIFF b/w • JPEG • JP2 colour • JP2 greyscale • Other	✓
MasterSize	Single line of text	Average file size of master image (in MB)	
Metadata	Choice	Choice: - MARC - MODS - MPEG - Dublin Core - MAB2 - Other	
Link	Hyperlink	URL to online version	
LibraryID	Single line of text	Library object identifier/catalogue ID (e.g. MARC digital object ID, dc:identifier)	

WP3selection	Yes/No	Whether this title is included in the WP3 evaluation dataset.	
TELPresentation	Choice	Choice: <ul style="list-style-type: none"> <li>- 1 – Full page view</li> <li>- 2 – Snippet view</li> <li>- 3 – Plain text view</li> <li>- 4 – Metadata only</li> <li>- 5 – Full page view (image server)</li> <li>- 6 – Snippet view (image server)</li> </ul>	
Created By	Person or Group	Automatically created per entry	
Modified By	Person or Group	Automatically created per entry	

Example of an entry:

ID	UID	LibraryID	Partner	NewspaperTitle	TELPresentation	WP3Selection
1	BnF_0001	FRBNF39294634	BnF	Le Journal des Débats politiques et littéraires	6 - Snippet view (image server)	FALSE
<b>StartDate</b>						
1814	<b>EndDate</b>					
1944	<b>Language</b>					
French	<b>FontType</b>					
Normal	<b>PageSize</b>					
Large	<b>Digitised</b>					
TRUE	<b>OCRed</b>					
TRUE	<b>Source</b>					
Original	<b>MasterFormat</b>					
TIFF greyscale	<b>MasterSize</b>					
34,55	<b>Metadata</b>					
MARC	<b>Refinement</b>					
OLR (CCS)	<b>Pages</b>					
198894	<b>Link</b>					
<a href="http://gallica.bnf.fr/ark:/12148/cb39294634r/date">http://gallica.bnf.fr/ark:/12148/cb39294634r/date</a>						

### 3.2 Data preparation

A number of preparatory steps need to be taken by libraries before they can have their data signed off for delivery to the refinement partners, and ingested into the automated refinement procedures.

The diagram below illustrates the workflow for data preparation by the libraries.

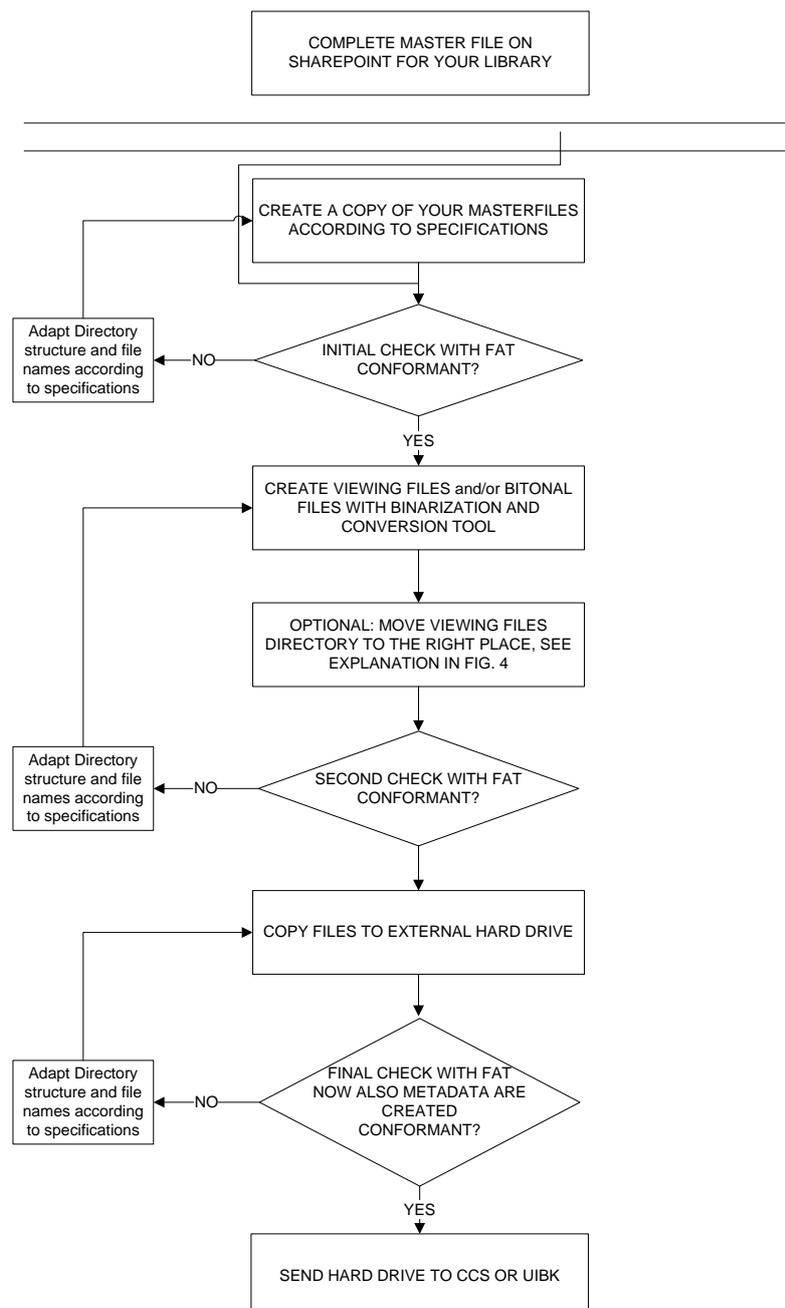


Figure 3: Workflow for data preparation

The libraries start with completing the master file on the project Sharepoint, as described in paragraph 3.1.1. This is the selection for refinement that will be sent to either UIBK or CCS. Each title has to be stored to the correct requirements for processing, as described in paragraph 3.3 below. When this is complete, the library can validate the files with the FAT (File Analyzer Tool). More information on this tool can be found in paragraph 3.4.

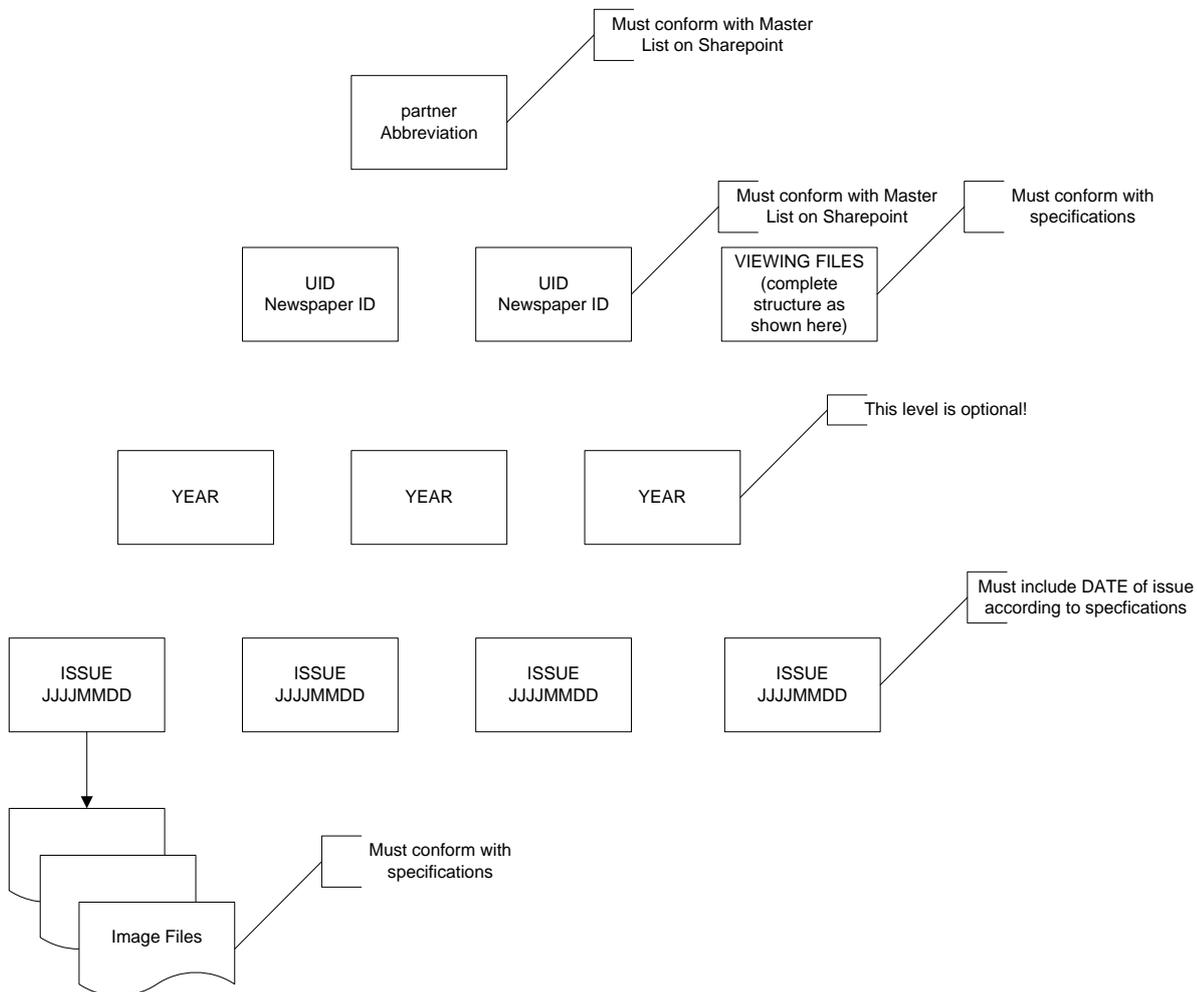
If the files are in concordance with the requirements, the library can continue to create bitonal files and viewing copies of the newspaper master images, for which the binarisation and conversion tools are provided, also described in paragraph 3.4 below. After this step, the files need to be checked with the FAT again to ensure the requirements are still met. If not, the library can adapt the directory structure accordingly. The viewing files should be stored in a similar matter as the bitonal master files. If not, the library can adjust the folder structure where necessary.

Once the files are stored correctly, they are copied to an external hard disk. For good measure, FAT is run again to do a final check on the requirements before the hard disk is sent to CCS or UIBK to do the processing.

### ***3.3 Delivery packaging***

Delivery from content holders to refinement partners needs to happen according to a strict system and schedule, so as to always guarantee a controlled processing of the approximately 10 million files that will be put through the refinement workflows altogether. The system is constituted by exact requirements for directory structure and file-naming in the delivery package that needs to be compiled on site by the libraries and then provided to the refinement partners accordingly.

Below diagram illustrates the exact structure of the delivery package and the according file naming requirements.



**Figure 4: Structure of delivery package and file naming requirements**

A strict file-naming system has been put into place to guarantee the tracking of the approximately 10 million files delivered from content providers throughout the whole processing. The FAT tool (cf. 3.4c) validates the data delivered by libraries against this file-naming schema. Any data that does not precisely follow this schema will be rejected and needs to be re-delivered according to the exact specifications by the library partner.

### 3.4 Tools

Altogether three software tools have been created by the University of Innsbruck to support the delivery process:

#### a) File Binarization and Conversion Tool (BCT)

This tool is used to binarise (= convert to black-and-white) master files and thereby significantly reduce the file size for the data transfer.

Internally BCT is using an algorithm from the National Centre for Scientific Research in Athens, called "[GPP](#)" (Gatos/Pratikakis/Perantonis). GPP is a state-of-the-art binarisation algorithm which is specifically optimised for OCR purposes. Tests conducted by the University of Innsbruck have shown that there is almost no tangible effect on the OCR accuracy when using bitonal image files binarised with GPP in comparison to greyscale or colour images, whereas through binarisation and

(lossless) compression with TIFF Group 4 the overall file size of the images can be reduced to less than 5% of the originals.

BCT also allows creating viewing images (= images with reduced resolution for display in a web browser) according to the specifications provided by TEL. For this, the tool makes use of the GraphicsMagick image processing system. GraphicsMagick is particularly optimised for robustness and high throughput scenarios. Testing has shown that by replacing ImageMagick with GraphicsMagick, the processing time for creating viewing copies could be reduced by about 50%.

The tool is supported by a user guide available from Sharepoint.

#### b) File Renaming Tool (FRT)

This tool is intended to support libraries in the renaming of their content files and the creation of a valid directory structure for data delivery.

The tool is supported by a user guide available from Sharepoint.

#### c) File Analyzer Tool (FAT)

This tool is used to check the conformity of the files against the specifications as well as to create metadata that are used for the OCR processing and for the creation of the final METS files which are delivered by CCS and UIBK.

FAT will give feedback about whether all files and folders meet all the requirements set out in this document. Files and folders that do not conform to these specifications won't be accepted. The library will have the ability to pick out/correct the rejected images and repeat the analysis. In some cases only warnings will be given so that the further processing is possible but it is clear to the user that further action is required at a later stage.

The tool is supported by a user guide available from Sharepoint.

All tools are available from the FTP server of the University of Innsbruck, and via the Europeana Newspaper project Sharepoint server respectively.

### **3.5 Delivery**

Delivery of files from libraries to refinement partners will be dealt with using external hard disks. The drives must have a connector for either USB 3.0 or eSATA protocol in order to ensure quick transfer times and should be formatted with the NTFS file system.

The hard disks contain complete newspaper titles, with the binarised images for OCR and the viewing copies packed together in a ZIP archive.

There will be a Sharepoint delivery tracking system for both the workflows of UIBK and CCS. The CCS delivery tracker system is based on the Delivery Manifest Schema CCS and will also make use of a tracking list for the hard disks that are being sent, based on the same system that has been proven in large-scale digitisation projects with customers such as for example the National Library of the Netherlands. A generic command line tool to update such and similar Sharepoint lists is also available and has kindly been circulated with the technical partners by CCS.

UIBK is currently using a Sharepoint list for progress monitoring, but is evaluating the CCS Delivery Manifest and command line tool and will switch to use it as well if applicable. There is also a specification available on Sharepoint for the UIBK delivery tracker, as well as an example of a delivery report.

Library	Delivery Date	Nr	OCR Start	Last Update	Nr Newspapers	Nr of Issues	Nr of Pages	Status
						<b>Sum= 196,783</b>	<b>Sum= 1,982,243</b>	
LFT	11/15/2012	1	1/7/2013	1/25/2013	15	93,220	828,747	OCR_Completed
LFT	12/11/2012	2	1/24/2013	1/29/2013	1	1,950	23,196	OCR_Completed

Status	Next Step	Newsp. completed	Issues completed	Pages completed	Error Report	% OCR Progress	DiscID
				<b>Sum= 1,843,164</b>			
OCR_Completed	Mets Creation	15	93,220	828,747	LFT1 Final report	100.0 %	5870CDFA
OCR_Completed	Mets Creation	1	1,950	23,196	LFT2 Final report	100.0 %	85C2414

**Figure 5: Example of the UIBK delivery tracker**

Delivery ID in DB:	1
Delivery Date:	08.11.2012 10:20:00
ROOT ID:	LFT
DISC ID:	5870CDFA
STATUS:	OCR_Processed
NR ISSUES:	93220
NR ISSUES COMPLETED:	93212 ( plus 8 issues were delivered empty)
NR PAGES COMPLETED:	828747
NR ISSUES WITH PROBLEMS:	0
-----	

**Figure 6: Example of a UIBK delivery report**

As the organisation of delivery of files from and to content providers is slightly differently organised at UIBK (who prefer receiving few but large chunks) than at CCS (who foresee smaller monthly deliveries), it is also possible that there will be two distinct tracking systems in the end. However, this does not cause any problems as there will always be a clear indication of current status available from the Work Package 2 calendar on the internal Sharepoint server.

## 4. Conclusion

Deciding on specifications for the refinement of 10 million newspaper pages, by three service providing institutions is a challenging task. However, the partners in the Europeana Newspapers Project have set up the optimal workflow for this undertaking and will combine efforts in creating a valuable new resource for the Europeana users.

The master list is where the process of refinement begins. The libraries indicate their newspapers there, with all relevant information concerning their newspaper titles. Consequently, they prepare their data according to the specifications with the help of the tools made available by the University of Innsbruck, before sending it via hard disk to either UIBK or CCS. The material gets processed accordingly (OCR/OLR and NER), while the libraries can keep track of the status via various tracking tools on the project Sharepoint.

Finally, the libraries receive their enriched newspapers for their own institution as zipped METS/ALTO packages according to the ENMAP profile defined in Work Package 5. UIBK and CCS also forward all material to Europeana, for the final publication via the newspaper content browser from Work Package 4, as to combine the 10 million refined newspapers pages into a unique dataset for the European public.