DELIVERABLE

Project Acronym: Europeana Newspapers

Grant Agreement number: 297380

Project Title: A Gateway to European Newspapers Online

D2.4 Recommendations on best practices for refinement of digitised newspapers in Europeana

---

Revision: 1.0

Authors: **Lotte Wilms, KB**

Günter Hackl, UIBK

Claus Gravenhorst, CCS

Contributions: **WP2 partners**

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

Revision History

Revision	Date	Author	Organisation	Description
0.1	26-01-2015	Lotte Wilms	KB	Outline
0.2	18-03-2015	Lotte Wilms, Claus Gravenhorst, Günther Hackl	KB, CCS, UIBK	First Draft
0.3	25-03-2015	Sandra Kobel	SBB	Textual Corrections
0.4	27-03-2015	Artus Zogla	NLL	Internal Review
0.5	31-03-2015	Lotte Wilms, Claus Gravenhorst, Günther Hackl	KB, CCS, UIBK	Updated
1.0	07-04-2015	Clemens Neudecker, Sandra Kobel	SBB	Final Version

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This deliverable contains original unpublished work except where clearly indicated otherwise.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. Management summary

Completing the refinement of over 10 million historical newspaper pages of 14 European partners within three years can be described as a major feat. During this process, decisions were made that were of impact to the workflow and outcome of the refinement and naturally, a great deal was learned. To ensure that other institutions wanting to refine digitised newspapers for their collection and that of Europeana can use the knowledge that was gathered in Europeana Newspapers, this deliverable can function as a baseline to begin their projects with a head start. Next to experiences with the refinement, the project also offers tools and software that are helpful in (pre-)processing of digitised historical newspapers.

This deliverable describes the workflow of the project, the lessons learned with regards to the technical aspects and planning within the project, and finally provides recommendations of the three technical partners in the project: University of Innsbruck, who was responsible for the Optical Character Recognition (OCR) of more than 8 million pages, CCS, who refined more than 2 million pages with Optical Layout Recognition (OLR), and the National Library of the Netherlands, who worked on Named Entity Recognition (NER).

With this overview of the refinement processes of Europeana Newspapers and the resulting recommendations, an insight into the experiences shall be provided and the deliverable shall serve as a starting point for new projects aiming at the refinement of historical newspapers for Europeana.

2. Refinement of digitised newspapers in Europeana Newspapers

2.1 Project workflow

Organising the refinement of over 10 million pages requires a very clear and highly standardised workflow to ensure that all parties are aware of their tasks at all times and can always track their data at any point in the defined workflow. The setup of this process is described in more detail in deliverable *D2.2 Specification of requirements of OCR and structural refinement-services for digitized newspapers in Europeana*, but a short recap will be given here in order to evaluate the current process and provide our recommendations for future projects.

2.1.1 Pre-processing

All libraries input the information about the data they selected for refinement into a central Master List on the project extranet. This list contains all necessary information for the refinement, such as language, image size, font type and total number of pages. Next to this, the list also contains the metadata about the object itself that will be shown in the image browser, such as title, date range and the identifier of the library.

Once the libraries completed the Master List, they prepared their files according to the specifications and tested their compliance with a specifically built tool, the File Analyzer Tool[1]. This tool checks the validity of the images, metadata and the standardised directory structure. If the files were checked by the FAT tool and no errors occurred, the libraries are safe to continue with the creation of viewing copies needed for the presentation and the conversion of the images to bitonal (i.e. black-and-white), using the File Binarisation and Conversion Tool[2]. This was done to reduce the file size of the images, thereby making it possible to ship the large amounts of data via regular cost-efficient hard disks. However, this is only an intermediate step in the technical refinement process. Should a library have scans

[1] More information about the FAT (File Analyzer Tool) tool can be found in deliverable D2.2.
[2] More information about the BCT (Binarisation and Conversion Tool) tool can be found in deliverable D2.2.

in greyscale or colour available, these will still be used instead for rendering the presentation for the end-user. The final pre-processing step at the libraries was the copying of the data to the hard disks and shipping it to the technical partners for the actual refinement.

2.1.2 OCR at UIBK

Within the Europeana Newspapers project, University of Innsbruck (UIBK) was the main provider for OCR. Around eight million newspaper pages were enriched with OCR by UIBK as part of the project.

The work was carried out within the “Abteilung für Digitalisierung und elektronische Archivierung” (Department for Digitisation and Digital Preservation”) who have gained extensive experience with OCR by participating in various European projects 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.

UIBK uses 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, UIBK modified 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 used FineReader SDK was at each time the most recent and robust release - at the beginning of the project v10 was available and after the first half of the project v11 was released and accordingly used.

In the first year of the project a detailed workflow was conceived, containing specifications and tools for easy delivering and supervised refinement. After year one the production cycle started. The lessons learned and best practice recommendations for the OCR at UIBK are summarised in the corresponding sections.

2.1.3 OLR at CCS

In addition to the roughly eight million pages of “regular” OCR provided to the project by the University of Innsbruck, another two million pages are OCRed with additional structural refinement, referred to as OLR (Optical Layout Recognition). In addition to merely

recognising the text, OLR includes advanced features such as the separation of articles and the classification of pages (e.g. into advertisements, titles pages, etc.).

The OLR workflow is run by CCS, a company specialising in newspaper digitisation and having ample experience in large-scale newspaper digitisation projects. CCS uses their in-house docWorks software technology for the project. This is an intelligent application for automatic conversion, structuring, and indexing of printed or electronic documents such as books, journals, newspapers and magazines. With docWorks, it is possible to locate and categorise key data from a variety of documents.

After raw data verification and ingest, the conversion process starts off with page analysis to determine page frames, 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.

Each automatic detection step can also be followed by manual verification, depending on the quality level required. For mass digitisation projects, manual verification is typically reduced to a minimum, but the content holders participating in the OLR workflow were provided with three alternative solutions for manual quality assurance by CCS, thus allowing them to at least sample the quality, perform some manual corrections and understand the importance of good raw material for optimal automatic results.

2.1.4 NER at KB

For a subset of the OCRed content from partners in Dutch and German language, the National Library of the Netherlands (KB) provides tools and technologies for the extraction of named entities such as person and place names or the names of organisations. This greatly improves the usability of the full-text for further text-mining and scientific purposes, since users can search for specific individuals or places. This also allows the cross-linking of the refined newspaper content with other online information resources such as authority files and the linked open data cloud.

The NER system that is implemented by KB for the project builds on prior work and experience derived from the IMPACT project. The Stanford University NER tagger, a mature and widely used machine learning tool for NER, is used and was also further extended for

the project. The software itself, including all necessary adjustments as well as the data that is used for training, is subsequently published under an open license[3].

Given that approximately 25% of the full text produced by the project is in the French language and as NER for French was not foreseen in the Description of Work, the National Library of France (BnF) submitted a proposal to the Project Management Board for the development of NER resources for French via a collaboration with the ACASA LIP6 Group of the Université Pierre et Marie Curie in Paris. An agreement was reached that ACASA LIP6 would develop technical resources for French NER in close collaboration with the BnF and the Europeana Newspapers project. These technical resources follow the design principles of the NER approach chosen by the project, and are made available under equal terms as for the other languages already supported.

The step following Named Entity Recognition is the disambiguation and linking of extracted named entities. Disambiguation of entities and referencing of authority files are more advanced ways to refine NER results, which in turn require a more sophisticated implementation, but at the same time allow for a much more useful presentation of the results. The KB received extra funding to work on this extra step in the workflow. The software which has been produced for this task is freely available under equal terms as the NER package[4].

A test set for Dutch, German and Austrian of around 1.000 pages was produced and evaluated by the KB, after which it was delivered to The European Library, who has used the material as a use case to test the integration of named entities into the main presentation interface. The French software was only delivered at a later stage and thus no tagged pages were produced for Europeana within the project lifetime. Given that all software is freely available, the libraries interested in processing their entire collection can do so without problem.

[3] See https://github.com/KBNLresearch/europeananp-ner and http://lab.kbresearch.nl/static/html/eunews.html
[4] See https://github.com/KBNLresearch/europeananp-dbpedia-disambiguation

2.2 Lessons learned

2.2.1 Technical

2.2.1.1 OCR

At the beginning of such a large scale project it is very hard to predict the various kinds and numbers of problems you run into. The only possible preparation is to devise a plan how to react and how to avoid possible technical problems. Working with checksums to observe the integrity of files is one example of such risk management. Introducing the File Analyzer Tool (FAT) during the first year was another. And this decision turned out to be absolutely right. To verify each file already at the library site prevents server stops and failures at the technical site and avoids overhead in exchanging files for re-processing later on. Also, collecting all the metadata simultaneously with the corresponding files is definitely recommended to have a smooth workflow.

Unknown problems clearly are not to be foreseen. For example, UIBK processed several newspaper titles for SBB and SUBHH containing very large stock market tables with tiny text (see Figure 1).

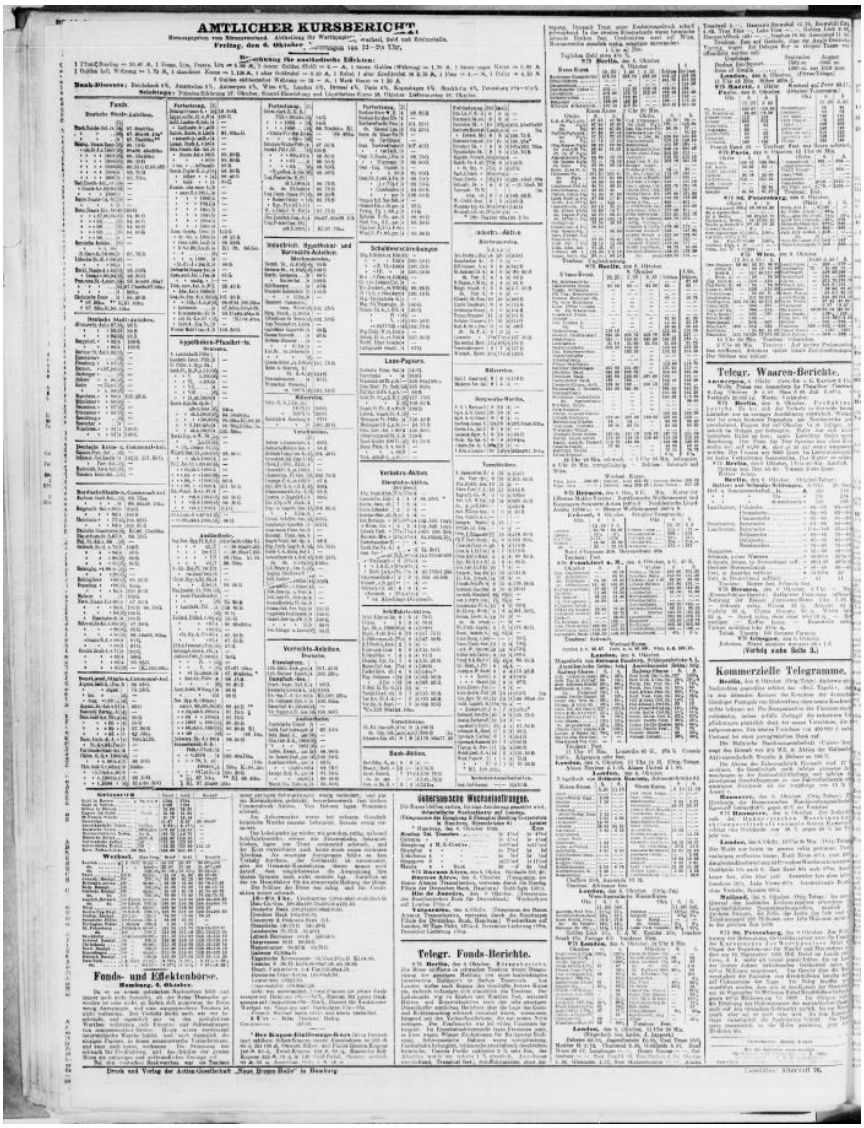

Figure 1: Example of stock market page

In the best case scenario, ABBYY FineReader simply computes these pages several times longer than normal pages, but in the worst case the engine stops and recognition fails. Here, UIBK came up with an extra strategy for such pages. The OCR engine offers a fast mode where recognition works 2.5 times faster but OCR quality decreases slightly. Other pages, not working with this setting, got recognized without table recognition – just the plain text. However, this meant extra work and as said before was not predictable at all. Here, the

recommendation would be to have a really good knowledge about the collection so that this information can be used for the OCR strategy planning and runtime calculation.

Beside the pure refinement of pages, the intention of the Europeana Newspapers project was also to bring up a standardised delivery package and this was the point where the refinement process was linked to another work package (WP5). The Europeana Newspapers ENMAP[5] format defined in WP5[6] was produced in WP2 for each single newspaper issue in a post processing step after text extraction had happened. All necessary technical, as well as descriptive metadata, were collected via the so called Master List and with help of the FAT tool as a first step. Due to the ongoing parallel ENMAP design, more metadata were collected than finally needed for production. However, it is better to collect more to ensure that nothing is missing than to have to ask for more information, which might slow down the process. Consequently, the updated and recommended Master List for future newspaper digitisation projects will be for that reason a somewhat simpler list and containing less fields.

The main tools developed to support WP2 (File Rename Tool, Binarization and Conversion Tool and File Analyzer Tool) were updated and refined several times during the project life time and are for that reason stable, with up-to-date downloads available as open source at GitHub[7]. Those tools can be recommended to use for future refinements.

2.2.1.2 OLR

Using a predefined list of prerequisites for file naming, folder structure and image format definition meant CCS encountered many fewer of the typical problems that normally crop up at the start of production. However, later changes to this list meant that corrections had to be manually adjusted. In order to overcome this extra work, it is thus recommended finalising the list before the production process or to implement a client interface for changes and subsequently taking any alterations into account in the planning.

Another weighty issue was the forecast of required Fraktur OCR pages with the licensing model used for this project. Each individual zone which was modified by the partner

[5] ENMAP – Europeana Newspapers Mets Alto Profile
[6] See D5.2 First public release with updated online resource for documentation
[7] See https://github.com/dea-uibk

institutions during their correction caused a re-OCR, which is counted by ABBYY as a single page – and a printed newspaper page can easily contain 50+ zones. In consequence, the final amount of licences required was much higher than expected. This is something that should be taken into account for future projects and could possibly lead to contracting different licensing conditions or the use of a different OCR engine.

Deciding beforehand what type of output format will be used for the whole process is of great importance, because making changes to the format can lead to differences in the final collection. When using a format that is newly developed or that will be developed within the lifetime of the project, it can be recommended to work with test batches in order to finalise the format before starting the bulk processing.

Finally, it is also important to take into account what the presentation portal will look like, as decisions made for the development of such a portal, like image specifications or format requirements, might be of influence to the digitisation workflow.

The offshore quality control performed by the partner institutions worked as expected, and no technical surprises or problems were encountered. It did take considerable time for some organisations to decide on which of the set-up options to use, and to get their infrastructural requirements worked out, but CCS was still able to complete all the different set-ups on schedule.

2.2.1.3 NER

One of the major lessons learned by the KB in the NER workflow was the decision to work with output from a previous project that appeared not to be production ready. This meant working with tools (especially the tool used to annotate data for machine learning) that required special training for the users and that were frequently updated. However, as tool production is not the core business of the tool provider, these updates were not done via a versioning system, but were provided via e-mail or other means of data transfer instead. This produced some unexpected issues with incompatibilities when exporting the data and the tool provider had to support the project in the correction of errors in the training data.

2.2.2 Planning

2.2.2.1 OCR

It is in principle very difficult to forecast and plan the average recognition time per page for such a mass refinement with several partner libraries and many different newspaper titles. Consequently, UIBK made predictions at the start of the projects, using a sample set concerning newspaper size, character number, text type and language. Already at this time it could be observed that UIBK had to deal with newspaper sources widely varying in characteristics and quality. For example, the French collection, of which the majority page format was A2, included pages that could be slightly dirty, which increased the number of characters on the page due to misrecognition (see Figure 2). Here, there average recognition time with the UIBK server infrastructure (4 servers with each 8 cores) was around 10 seconds.

LES

PROPOS DU BOULEVARD

M. Poubelle, préfet de la Seine, recevra, aujourd'hui, à deux heures, en son cabinet, à l'Hôtel de Ville, tous les fonctionnaires municipaux. Le défilé commencera par le président et les membres du conseil, de préfecture, et se terminera à cinq heures par le colonel commandant le régiment des sapeurs-pompiers et son corps d'officiers.

Le bureau des longitudes est ainsi constitué pour l'année 1886 : MM. Rideau et Ossian Bonnet ; astronomes : MM. Faye, Lœvy, Janssen et Tisserand ; membres du département de la marine : MM. les amiraux Paris, Mouchez et Cloué ; membres du département de la guerre : M. Perrier ; géographe : M. d'Abbadie.

La Société des artistes français vient d'arrêter le règlement du Salon de 1886.

Rien n'est changé quant à la date d'ouverture, mais certaines modifications ont été apportées au règlement en ce qui concerne le dépôt des œuvres.

Le délai pour la section de peinture ne sera plus que de quatre jours, au lieu de neuf.

Cinq jours seront accordés pour le dépôt des œuvres de sculpture, et enfin, trois jours seulement pour les ouvrages de gravure et de lithographie, ainsi que

vulgairement appelés gros bonnets.

Agréez, etc., etc.

Un de vos lecteurs.

La Société Hongroise de Paris donnait hier, dans les salons du Grand-Véfour, sa soirée musicale et dansante annuelle.

Un incident, qui ne figurait pas au programme, a marqué cette fête de famille.

A dix heures, la délégation française, M. de Lesseps en tête, a fait son entrée dans les salons. Elle venait, en souvenir de la magnifique hospitalité qu'elle a reçue en Hongrie, porter à ses hôtes nos grois ses vœux pour l'année nouvelle.

Reconnu dans le nombre : MM. Armand Gouzien, le colonel Lichstenstein, Delibes, Massenet, Clairin, Félicien Rops, Escalier, Maurice Bernard, Louis Ulbach, Emile Blavet, les docteurs Robin et Pozzi, Lebrasseur, Mario Proth, Aubey, Louis Ratisbonne, E. Yung, Abraham Dreyfus, etc., etc.

M. Ad. de Berthal, le distingué compositeur, a souhaité la bienvenue à la Délégation dont M. de Lesseps s'est fait ensuite l'interprète dans un de ces speechs vibrants et chaleureux dont il a le secret.

Le traitement des académiciens n'a rien de bien tentant et n'était l'immortalité qui lui est adjointe, la place ne serait guère enviable.

L'Académie française, l'Académie des inscriptions et belles-lettres, l'Académie des beaux-arts, l'Académie des sciences morales et politiques comptent chacune

à cet homm
déste.

Le Conseil
séance, a vo
professionnel
cieux et artis

NOUV

T... est un
sédé cent sou
Il avait ces
— Vous pa
dit un de ses
— Je suis
quelque argei
accourent en
ne dors plus,
— Voilà c
joie !
— Certain
les infirmités
Voilà les une

Un pauvre
faire une sou
trop tôt, car
bouillon aux
an.
Mais il ava
moins possib
tane trop étro
nures.
Un jour, il
— Ça ne p
dit-il, je ne p
la messe, qua
le diable !

Figure 2: Example of a dirty page

This means that with only that kind of pages UIBK would have needed over 80 million seconds for refinement, exceeding the available time frame of 2 years by over 25 %. So the number of characters on a page is one important calculation factor for the planning of any digitisation project, while the quality of the page is the other.

However, more factors have to be considered. If one page contains Antiqua, as well as Gothic text type, ABBYY FineReader can handle both in the same recognition process automatically, but needs more time to do this. The same is valid for more than one language. Each additional language degrades the refinement speed a little bit. Having more than 4 languages at the same time slows down the process significantly. In addition, during the start-up phase of a digitisation project, the decision should be made if pages get rotated automatically by the OCR engine or not. If the answer is yes than the consequence is that the OCR coordinates of the rotated page either do not match to the original page coordinates or that the rotated page must afterwards be regarded as a new 'original'. Also splitting double pages produces new facts. These decisions often trace a rat-tail of actions and errors and were therefore omitted for this project, while for smaller projects this could be valuable.

All the above points have to be taken into account for a precise OCR planning.

The time estimate can be done automatically with a subset of the designated OCR material. Moreover the quality and time estimate for the planned digitisation project can be combined. The initial subset must not be too big but must be representative and processed as in the later production OCR workflow. This way the partners get an impression of the OCR quality and the time needed for refining the whole batch without (time-consuming) counting of the number of characters or other similar estimations.

In this vein 'noisy' images could stand out and a pre-processing step can be used to filter out the background noise. Always trying to filter out would be too time expensive for mass digitisation projects like Europeana Newspapers with several partners and different sources.

As the planning phase showed, historic newspapers in some countries have small page size. The titles from the Austrian National Library (ONB) as well as those from Landesbibliothek Friedrich Tessman (LFT) titles are mostly A4 and were processed really fast, with less than 2 seconds per page.

Special cases were the titles of the Turkish partner, National Library of Turkey (NLT). NLT provided half of their pages with Latin text type but the other half containing Ottoman letters. Recognition for the latter one is not yet supported by the OCR Software and trying to process the pages with '*Arabic'* produced only poor output with only 20 % recognition rate. But the good news was that the layout recognition was quite successful and useable. The second half of NLT's delivery was therefore done to gather just the text layout, enabling them to prepare these pages for manual correction.

But calculating time in such a project is one aspect of planning, as is disc space. If using the original images for delivery, UIBK would have needed over 300 Terabyte storage space. This means - by using 3 TB discs - sending around over 100 discs during the project life time (assuming that the full storage quantity always gets used, which was not the practical experience) and a very expensive network attached storage on the refinement site as well as the presentation site would have been necessary. Therefore, the procedure was devised to binarise the images in a pre-processing step with just a small quality loss[8] with a storage demand reduction rate of almost 90%. This way UIBK have also reduced the manual work during the disc delivery. For presentation on the newspaper browser developed in WP4, JP2000 files were created from the original images.

On the library site, preparing a delivery means copying and may also mean digitising sources, binarising images, creating viewing files (both tasks are possible with the developed and provided BCT tool), checking file integrity (with the developed and provided FAT tool) and collecting metadata for further refinement (FAT supports this task as well). On the technical site, handling a delivery means communicating with the library, importing the delivered data to the server storage, importing metadata to the refinement database, starting and observing the OCR process, later on checking and possibly reprocessing any failed pages, starting METS[9] creation and delivering the refined pages back to the libraries and to TEL for presentation in the newspaper browser. In order to avoid too much manual work, few deliveries are better than many. But this was not always practical since preparing (digitising) batches for refinement were very time intensive and to avoid dwell time for the servers, smaller deliveries were also accepted. Deliveries are hard to plan in general, so one lesson

[8] See D3.5 Performance evaluation report
[9] METS – Metadata Encoding and Transmission Standard, http://www.loc.gov/standards/mets/.

learned was to build up a queue of recognizable newspaper titles in advance. UIBK also used titles which were not initially planned for refinement to use the full capacity of the servers at all times, e.g. ONB was able to deliver almost 3 million additional pages. Moreover, also associate partners were invited to deliver pages and so received refinement results for over 100.000 pages for the National Library of Luxembourg, around 10.000 pages for the National and University Library of Serbia and also the colleagues from the *Europeana Collections 1914 – 1918*[10] project were happy to OCR almost 300.000 pages within the Europeana Newspapers project. So all in all more pages than written in the DoW were refined but some shifting to the end of the project occurred since not all newspapers were already digitised at the beginning and scanning from scratch appeared to be an unexpected great deal of work for some libraries. To conclude, building up enough resources to process in time should also be a lesson learned.

The overall conclusion at the end of this project is that the planning and preparation phase of the first year, including the development of the helping tools, was very essential and undeniable important to reach the really ambitious work plan.

2.2.2.2 OLR

In all mass digitisation projects preparation is a key, as even tiny deviations from the set plan can have major consequences and cause great delays with substantial rework requirements. At the same time, a certain degree of flexibility is needed to deal with the particularities that occur in any project. The best method to cover both these demands is to provide as much transparency as possible right from the very start.

CCS achieves this by using a SharePoint platform for all its projects. For the Europeana Newspapers project, a main SharePoint site was created, with access being granted to all five OLR partner institutions as well as the project management level. This SharePoint contained a number of tracking lists, important announcement sections as well as a common repository for project documents such as the project plan. A sub-SharePoint for every partner institution, accessible only for that particular institution (and again of course project

[10] See http://www.europeana-collections-1914-1918.eu/

management), was also created, allowing partners to address their individual issues that were of no interest to the other parties, to CCS.

One of the key features on the main SharePoint was the delivery list that gets automatic additions for every delivery created in the CCS system. Thus every partner was able to track progress on their items and feel confident that all their material has been processed as it should be, and CCS recommends the use of such a list for future projects.

The site also contained an HDD tracking list. This is used to make sure the whereabouts of every hard drive are tracked. Having five different institutions, some with multiple deliveries of the same material (which will be touched upon further in this paragraph), means CCS needed to make sure that only the correct data was imported, and that all the hard drives were returned to their proper owners at the end of the project. Having the information on SharePoint meant any ambiguities could be resolved quickly and easily.

As the docWorks software provides an opportunity for manual correction of the automatically detected results, the partner institutions were given a choice of three options to perform not just a quality review of the data, but also to perform manual corrections on those results. The options consisted of

a) having the conversion happen at the institution itself (chosen by the National Library of Finland as they are long-time users of docWorks)

b) having the conversion happen at the CCS offices in Hamburg, and the final QA by the libraries taking place via internet transfer (so-called RemoteQA solution; chosen by the National Library of Estonia, Berlin State Library and State and University Library Hamburg)

c) having the conversion happen at the CCS offices in Hamburg, and the final QA by the libraries place by hard drive shipments of the processed data to the library for their review (not chosen by any partners and not recommended by CCS because of the large administrative overhead involved)

A detailed explanation of the three scenarios, the system requirements and the possible QA steps were provided in individual teleconferences with each partner, and followed up by the

provision of PowerPoint presentations and checklists that were added to the SharePoint for easy reference.

In addition to these three originally proposed options, it was revealed after lengthy discussions that the high security levels at the National Library of France (BnF) would not allow the remote installation of docWorks on a BnF machine, and so a fourth solution was needed. CCS created a virtual machine for the BnF, allowing them to access their data for review and QA by connecting to CCS's systems.

Despite the detailed and lengthy preparatory work, emphasis on the quality of the material having a direct impact on the quality of the results, and individual training sessions for each partner on how to use docWorks for their QA, the time and staff needed for the implementation of the technical set-up as well as the manual review and/or correction of data was underestimated by almost all partners, leading to delays both at the beginning of the project as well as along the way. Enough time should be allowed for these phases in future projects if the institutions are using software they have no previous experience with, or the degree of QA should be adapted to achieve realistic timelines.

Again, as with the OCR process at UIBK, CCS also encountered issues with data delivery. Next to this, there was also an issue with one library where the material that was selected appeared to be unsuitable after the initial processing, which meant that the library had to reselect images and process those with the ENP workflow, delaying the original schedule. CCS would recommend allocating sufficient time for material selection and extraction by the libraries to make a thoroughly checked selection.

The build-up of pages in the final months was also increased by the need for additional ABBYY OCR Gothic licences. The original calculation was that 2 licence pages would be required for each image page. This was revealed to be a drastic underestimation, as not only did the physical page sizes vary greatly, but docWorks also required additional licence usage because individual zones and not just entire pages were OCRed in the workflow. Depending on the type of OCR engine, it is important to check the number of OCR requests per page in the initial phase of a project. This is an important conclusion and should be considered for future projects.

Despite mitigation strategies and contingency plans being in place at CCS to handle risks, some events could not be foreseen and caused additional efforts, such as a complete system crash at one of the institutions, which meant loss of performed corrections, a second set-up, recovery actions etc.

While CCS is always happy to customize workflows to the needs of its customers, implementing individual solutions for various partners in the Europeana Newspapers project meant that the concept of mass digitisation had in some cases to be adapted. Examples are additional QA steps to perform page sequence corrections that were spotted at a late stage or correction or wrong information in the Master List regarding the fonts.

Again, it is important to allocate sufficient time for the preparatory stages. The better the preparation of the material, the better the automatic results and the lower the number of in-process changes required, which translates into fewer delays.

2.2.2.3 NER

Given that the process for Named Entity Recognition can only be done once OCR is available, the first months of the project were spent researching and evaluating existing tools and setting up the workflow for the project. Since the KB was already in possession of digitised newspapers, these were used for this purpose. Once material from the Landesbibliothek Friedrich Tessman (LFT) and the Austrian National Library (ONB) became available, they could start the training process of their material. However, due to some difficulties with the export function of the training tool, this process took more time than anticipated and a reset of the underlying database was needed on a regular basis, again slowing down the training process. These problems are solved in the latest release of the tool. Once the training was done, the material was exported and corrected and the KB could use it to produce the language models and finalise the software. Each library then provided the KB with a small test set of around 1.000 pages that was processed and delivered to both the libraries and The European Library for inclusion in the search portal.

Ultimately, the planning for the NER process was adequate and would the KB not have had personnel issues, the NER would have been finished within the original project lifetime.

When looking at the planning, it is most important to take enough time to evaluate the output of the software and for adjustments where necessary in order to get the best results.

2.2.3 Evaluation

2.2.3.1 OCR and OLR

The evaluation of the OCR produced by University of Innsbruck and CCS has been done by the University of Salford in WP3 and is described in *D3.5 Performance evaluation report*. The text below summarises this deliverable.

> In general it can be concluded that the produced results, especially with regard to the overall text accuracy, are of good quality and fit for use in a number of use scenarios. Moreover, technical decisions that were made during the setup of the production workflow could be confirmed. A number of observations (e.g. on the recognition performance for certain languages and particular layout problems) show mainly the limitations of current state-of-the-art methods rather than issues with the implemented workflow. In terms of layout analysis capabilities there is still room for improvement and any progress in this area could have a great impact on the usefulness of OCR results for more sophisticated use scenarios.

2.2.3.2 NER

Named Entity Recognition is typically evaluated by means of Precision/Recall[11] and F-measure[12]. Precision gives an account of how many of the named entities that the software found are in fact named entities of the *correct* type, while Recall states how many of the total *amount* of named entities present have been detected by the software. The F-measure then combines both scores into a weighted average between 0 – 1.

Each collection was evaluated separately by the KB:

[11] See http://en.wikipedia.org/wiki/Precision_and_recall
[12] See http://en.wikipedia.org/wiki/F1_score

Dutch	Persons	Locations	Organizations
Precision	0.940	0.950	0.942
Recall	0.588	0.760	0.559
F-measure	0.689	0.838	0.671

These figures have been derived from a 4-fold cross-evaluation[13] of 25 out of 100 manually tagged pages of Dutch newspapers from the KB. The results confirm the fact that the Stanford NER tagger tends to be a bit "conservative", i.e. it has a somewhat lower recall for the benefit of higher precision, which is also what was aimed for, as this is the most valuable for our users.

There were fewer pages available for evaluation for German and Austrian, due to the problems with the export function of the training tool that resulted in several pages being not useable. The decision was made to use as many as possible for training, which resulted in a smaller evaluation set. Therefore, the outcomes are not split up per category, as this would provide too little entities for a good evaluation. Five pages from LFT and six pages from the ONB were used for the following evaluation.

German	LFT (German)	ONB (Austrian)
Precision	0.76	0.79
Recall	0.69	0.66
F-Measure	0.72	0.72

[13] See http://en.wikipedia.org/wiki/Cross-validation_(statistics)#K-fold_cross-validation

The French material was evaluated by LIP6, which resulted in the following figures:

French	Persons	Locations	Organizations
Precision	0.84	0.80	0.70
Recall	0.83	0.78	0.49
F-measure	0.83	0.79	0.59

3. Best practice guidelines for OCR

3.1 Preparation

As noted already in the OCR Planning chapter it is very important to have a detailed understanding of the sources to reach the best refinement results. This means that the preparation time should not be underestimated and already in this phase 'Evaluation' can help to foresee the error rate of the material. This means having raw material which is comparable to one in the evaluation dataset the OCR results are supposedly predictable to a certain degree. Moreover, this means that predicting very bad OCR results could lead to the decision of rescanning. Clearly, this is not practical for material which is not available any more. Some newspapers of one of the partner's library were destroyed by fire and therefore the raw material is as it is. The only thing that can be thought of then is to do OCR correction as post processing step. Promising automatic correction tools exist but there cannot be expected too much. This is the reason why sometimes scanning from scratch would be the better long term strategy in some cases. The recommendation would thus be to collect a representative dataset out of the raw material, do some tests, evaluate, and only then continue.

At the very beginning of the OCR process flattening and cropping the images is to be suggested. Moreover the file and folder names and folder structure should meet the specifications in *D-2.2 Specification of requirements of OCR and structural refinement-services for digitised newspapers in Europeana*. The more time content providers invest in these enumerated preparation steps, the less negative surprises they will have.

Very important is to have the publishing date as part of the issue name to allow a day search in the final presenting system as this is the most helpful search filter later on. It was quiet time-consuming for at least one ENP-partner to find and write the date to each issue folder. Ten people were involved for over one month to prepare and mainly rename the folders to achieve the defined specifications. Therefore the File Rename Tool (FRT) was developed which helps to reorganize the file and folder structure when needed in a semi-automated way but needs to be adapted for each single case.

For the delivery preparation the workflow of this project is recommended. The File Analyzer Tool helps to verify the file and folder structure, checks the validity of the files and thereby avoids refinement interruptions later on and assists during entering the language and text type as well as delivering other metadata for the METS creation at the post processing. Moreover, the FAT calculates checksums for file integrity verification. The output of the tool gets delivered to the technical partner together with the pages for further refinement. Once one gets used to this - in the meanwhile approved and standardized - workflow it stays each time the same.

Creating bitonal images is recommended when dealing with mass refinement like in this project. In a small scale project the original files can act as input files for the OCR software as well.

3.2 Refinement

For the refinement a batch tracker list was introduced on the Europeana Newspapers Sharepoint site. This way, the library was always informed about the status of their actual deliveries and at the same time both sides saved lots of communication time.

For internal monitoring a database was used. Each delivery, each newspaper, each newspaper issue, and each single newspaper page was represented in this database. All necessary refinement steps (OCR, failure handling, METS creation) used this database as checkpoint and communication interface. Especially for failure handling it is important to find all uncompleted pages in the first run, but the OCR process and METS creation also need actual and central information. This way the refinement process was traceable all the time.

During refinement it was useful to do sample checking, e.g. check one file out of 1000. This way a wrong setting or some other error can easily be found and refinement can be repeated with no big time loss. This was a further quality assurance effectively tested during this project and is to be recommended for further projects.

3.3 Post-processing

As already mentioned, but out of scope for this project, are post OCR correction tools that can help to improve OCR quality. An interesting approach is that of the CIS-LMU Post Correction tool[14].

The definition and usage of the Europeana Newspapers ENMAP package was in the focus of this project. For the creation of the METS all collected metadata was imported into the database and written to the output METS during the post processing. METS/ALTO[15] seems to be a proper suggestion since it is commonly used in newspaper digitisation projects, e.g. Chronicling America[16] from the Library of Congress, Trove[17] from the National Library of Australia, Papers of Princeton[18] from Princeton University Library to just name a few. The advantages are clear: a highly standardised output format helps exchanging and using the results among several partners and guarantees long-term sustainability too. Additionally someone can easily create other formats from METS/ALTO or search for tools which can handle this widely spread METS/ALTO format.

[14] See http://www.digitisation.eu/tools-resources/tools-for-text-digitisation/cis-lmu-post-correction-tool-pocoto/
[15] METS – Metadata Encoding and Transmission Standard; ALTO – Analyzed Layout and Text Object
[16] See http://chroniclingamerica.loc.gov/
[17] See http://trove.nla.gov.au/
[18] http://theprince.princeton.edu/princetonperiodicals/cgi-bin/princetonperiodicals

4. Best practice guidelines for OLR

4.1 Preparation

Comprehensive planning and detailed project specification are vital elements of successful in-house or outsourced OLR projects. For good communication and effective project control, dedicated project managers with clear responsibilities should be in place. A collaboration platform should be foreseen to enable fast and transparent communication. At least for mid- and large-scale projects this is a must.

CCS recommends the use of a distributed workflow for OLR. Depending on the project and available resources, there are several options for setting up the workflow, offering the flexibility to either work fully in-house, fully outsourced or distribute related processing steps among institutions and service providers. To save human resources, automated processing can be performed in-house, while manual quality assurance can be outsourced to a service provider. For the in-house and distributed workflow options, having the technology in-house enables project control and verification of the quality of the output data. The fully outsourced solution works without any form of distributed workflow technology, and so should at least have applications in place to check the quality of the delivered and received data.

To ensure the best possible image quality, great care needs to be taken in the creation and selection of page images. If storage space and the data transfer scenario are not a limiting factor, greyscale and colour images should be used as input for the OLR workflow to create the best possible output quality. Images and metadata on issue level should be provided in a defined and standardised way. It is recommended to follow the related Europeana Newspapers recommendation.

As OLR enables logical structuring of newspapers down to article level, it is important to specify the tagging policy, thus defining page types as well as the structural elements like sections and articles with related elements such as author, headline, abstract, paragraphs, illustrations, captions, etc.

Deciding which metadata standards will be used to describe the digital object with technical, administrative and structural metadata is another critical factor. These standards should be

XML-based and enable the metadata to be ingested into the institutional repository and presentation systems.

4.2 Refinement

CCS uses Microsoft SharePoint as its collaboration platform. As much information as possible is shared via that platform such as delivery lists, details about redeliveries, clarification of open issues, hard disk drive tracking list for traceability of shipped items as well as shared documents like additional specifications, meeting notes, change requests, etc. This way, content providers are always informed about the status of their batches.

For the OLR workflow and the 2 million pages processed for 5 content providers, CCS chose to work from the best image quality available, either greyscale or colour. OCR as well as layout and structure analysis benefitted from built-in specialised processing algorithms to improve recognition rates. Whenever possible, using greyscale or colour images for OLR processing is recommended.

The use of a distributed workflow model meant that the content providers were able to gain experience and know-how about the digitisation and conversion process during the review and QA step of the workflow, which not only helped during the project but will also come in useful for future projects. Although reduced and limited, the amount of human resources required for this by the content providers should not be underestimated.

Naturally, in addition to the automated processing, CCS also performed a basic quality assurance to control and optimize the output quality.

The Europeana Newspapers project selected METS/ALTO as the open XML-based metadata standard to describe the digital object. This standard is maintained by the Library of Congress and widely used in the cultural heritage community. The highly standardised ENMAP[19], defined during the project, enables long-term preservation as well as data exchange and interoperability.

[19] ENMAP – Europeana Newspapers Mets Alto Profile

4.3 Post-processing

Although not in the focus of the Europeana Newspapers project, the OLR workflow enables reprocessing of already digitised and converted material to improve the quality of digital newspaper content in terms of adding information like page classes, classification of graphical elements like advertisements, photos, etc. and logical entities such as sections and articles. Already existing data like metadata, zoning information and OCR text can be reused and will lower the effort for reprocessing. Digital content available in ENMAP format can be ingested and re-processed. This enables content holders to select digitised collections or specific titles from their digital repository and reprocess those to improve access and distribution capabilities. In other words, existing page level data in ENMAP format can be converted to article level data and text correction can be performed on certain text objects like article headlines and illustration captions.

Text correction of the whole body of text is very time consuming and expensive. Therefore it is usually not part of mid or large-scale newspaper digitisation projects. Some projects offer a user text correction functionality as part of their web-based newspaper portals. Prominent examples are *National Library of Australia - Trove*[20] and *University California Riverside - CDNC*[21]

Apart from the above mentioned post-processing steps, the standardised METS/ALTO output can be used or transformed to feed other post-processing technologies such as Named Entity Recognition or, if we are dealing with article level data, other linguistics based technologies such as automated article classification and clustering.

And finally, an aspect of increasing importance is that the METS/ALTO output can easily be transformed to virtually any data format needed to use the data with current and future media devices such as tablets, touchscreens and mobile devices.

[20] See http://trove.nla.gov.au/newspaper
[21] See http://cdnc.ucr.edu/cgi-bin/cdnc

5. Best practice guidelines for NER

5.1 Preparation

These recommendations are intended for projects or libraries wishing to use the material created by the Europeana Newspapers project, i.e. NER for Dutch, German or French. When processing other languages, the software needs to be adapted to that particular language. It is recommended to get in touch with a party who is experienced in working with NER for your language to discuss the possibilities.

When preparing to use NER in your workflow, there are several choices that have to be made that can be of influence on the material that will be tagged. Firstly, it is important to select a dataset where the OCR quality is of a relatively high level. It is recommended to use only those newspapers where the OCR has at least a 70% accuracy rate on word level. Next to this, the selected dataset should ideally consist of newspapers that are uniform in language as this decreases the amount of training needed for the software.

5.2 Refinement

Before actually tagging the named entities, it is important that the software is adjusted to the material. Select around 100 pages in ALTO that are representative of the corpus and run them through the ENP NER Annotator[22] for pre-tagging. These results are loaded into the Attestation Tool[23] used for training and tagging each named entity in the training corpus. The files are exported in ALTO format and converted to the BIO format using the script that is delivered with the NER package by the KB. This step parses the tagged ALTO files into the BIO format that NER software uses as input. Such a file could look like:

In POS O
Nederland POS B-LOC
staat POS O
een POS O

[22] See https://github.com/KBNLresearch/europeananp-ner
[23] See https://github.com/INL/NERT

huis POS o

van POS O

Beatrix POS B-PER

Optionally a filter can be used to only use sentences with at least three named entities to increase accuracy and to reduce noise in the BIO files. The result of this step could be evaluated after processing the material. With the produced BIO files, the NER software can be trained to the specific material. Before training the software and processing the entire collection, selection of a small number of pages from the manually tagged collection should be done, that can be used to evaluate the results against and keep these files separate from the training set. Next to the BIO files, the NER software also takes gazetteers as input for the training. These are lists of named entities with their tag attached, such as:

Amsterdam LOC

Rotterdam LOC

‘s-Gravenhage LOC

Den Haag LOC

Groningen LOC

Eindhoven LOC

These can be specific to the location and specific wishes. For example, for the dataset of the Friedrich Tessman Library, the KB used a gazetteer that contained locations, names and organisations that were specific to the South-Tyrol area.

With the BIO files and the gazetteers, the NER software can be used to create the classifier after which the selected pages can be processed. Then, using the selected files for evaluation, evaluate the results using precision/recall[24] and if needed, repeat the workflow with for example more trained pages or different gazetteers to achieve the optimal results.

[24] See http://en.wikipedia.org/wiki/Precision_and_recall

5.3 Post-processing

Once one is happy with the results from the NE process and wishes to use the named entities for linking, it is important to disambiguate the results in order to increase the fidelity of the links. For this, KB has created a disambiguation service, which is available on Github[25]. To use this service, next to the tagged ALTO files, one would also need a DBPedia dump[26] and an installation of SOLR[27]. After having installed the service and used it to process the DBPedia dump one can then run the tool on the tagged pages and the tool will add a URI per linked named entity where applicable. This then allows linking the set to other sources, such as Freebase[28] or VIAF[29].

[25] See https://github.com/KBNLresearch/europeananp-dbpedia-disambiguation
[26] For example, available here: http://wiki.dbpedia.org/Downloads2014
[27] See http://lucene.apache.org/solr/
[28] See https://www.freebase.com/
[29] See http://viaf.org/

6. Conclusion

Over the past three years, the technical partners of Europeana Newspapers have worked very hard to refine over 10 million digitised historical newspapers. During this process, a great deal was learned about handling such a diverse collection of 14 European libraries, with over 20 languages. This deliverable combines the experiences and recommendations of working with OCR, OLR or NER for a large and varied collection of historic newspapers, with the intention to serve as input for other refinement projects focusing on historical newspapers for Europeana.

Each refinement stream firstly introduced the workflow that was used in the project, the lessons learned of working with said workflow with regards to the technical aspects, planning and also how these outcomes were evaluated. One of the major recommendations that all three partners give is to finalise as much of the workflow as one can before starting the major refinement. Changes to metadata, software, formats or requirements eventually lead to differences in the dataset and luckily, these differences were all manageable in Europeana Newspapers, but could prove to be quite a problem and should be avoided if possible. Finally, each technical partner (UIBK, CCS and KB), has described the ideal process for OCR, OLR or NER for digitised historical newspapers for Europeana, taking into account the experiences from the project and thus providing an overview of best practices for future projects.