

ECP-2006-DILI-510003

TELplus

Survey of existing OCR practices and recommendations for more efficient work

Deliverable number	<i>D1.2</i>
Dissemination level	<i>Public</i>
Delivery date	<i>31 July 2008</i>
Status	<i>Final</i>
Author(s)	<i>Joachim Korb, Austrian National Library</i>



eContentplus

This project is funded under the *eContentplus* programme¹,
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

¹ OJ L 79, 24.3.2005, p. 1.

Executive Summary	3
1 About this Document	4
1.1 Purpose of this document and its relationship with other TELplus documents.....	4
1.2 Input for this document.....	5
2 Survey of existing OCR approaches.....	6
2.1 OCR pricing	7
2.2 OCR software	9
2.3 OCR output formats.....	10
2.4 OCR quality assurance.....	12
2.5 Digital Library Systems	13
2.6 Search solutions	15
2.7 Access points	15
2.8 Summary.....	16
3 Good practice recommendations.....	18
3.1 Project purpose	18
3.2 Project size.....	19
3.3 Choice of material for OCR.....	20
3.3.1 Large scale projects	21
3.3.2 Smaller projects	22
3.3.3 OCR input.....	23
3.3.4 Digitised microfilm material as input for OCR	23
3.3.5 Avoiding duplication of work?	24
3.4 Pre-processing.....	24
3.5 Choice of OCR output format.....	25
3.6 Measuring OCR output correctness.....	26
3.6.1 Software log analysis vs. human eye spot test.....	26
3.6.2 Letter count vs. word count	27
3.6.3 Re-consider checking OCR output accuracy	28
3.7 Post-processing	28
3.7.1 Lexical correction	28
3.7.2 Collaborative correction	30
3.8 Cooperation.....	33
3.8.1 Inter-library cooperation.....	33
3.8.2 Private-public partnerships	34
3.9 Summary.....	35
References.....	37

Executive Summary

This document is one of five documents that will be published as part of TELplus workpackage "Making Searchable Digitised Images via OCR" (WP1) over the course of the runtime of the TELplus project. Three of these documents, the "Package of Specifications and Implementation Plans", the "First Set of Consolidated OCR Progress Reports" and the "Second Set of Consolidated OCR Progress Reports" (TELplus deliverables D 1.3, 1.4 and 1.5), are internal deliverables that follow the progress of the OCR work of the 14 content providing partners of WP1.

The other two documents, the "Survey of Availability of Digitised Images for OCR" and the "Survey of Existing OCR Practices and Recommendations for More Efficient Work" (TELplus deliverables D 1.1 and D 1.2), are public deliverables whose purpose it is to provide access to the findings of this workpackage to the wider public as well as to the TELplus partners.

As the title suggests, the "Survey of Existing OCR Practices and Recommendations for More Efficient Work" is designed to be two documents in one. The "Survey of Existing OCR Practices" is intended to supplement the "Survey of Availability of Digitised Images for OCR" in order to give a good overview of about the linguistic, geographic, and temporal scope of the material that is being OCRed for TELplus, about the different workflows that have been established at the partner libraries, and about the way the material is accessible from the libraries' websites.

The "Recommendations for More Efficient Work" are intended to give an overview of the points that have to be considered when planning an OCR project. They are based on the experiences of the TELplus partners and on information collected over the course of the project. They are intended less as a checklist to be ticked off when planning an OCR project, but more as a list of points to be aware of and of suggestions of how to deal with these points, once they become important for a project.

1 About this Document

1.1 Purpose of this document and its relationship with other TELplus documents

This document is one of five documents that will be published as part of TELplus workpackage "Making Searchable Digitised Images via OCR" (WP1) over the course of the runtime of the TELplus project. Three of these documents, the "Package of Specifications and Implementation Plans", the "First Set of Consolidated OCR Progress Reports" and the "Second Set of Consolidated OCR Progress Reports" (TELplus deliverables D 1.3, 1.4 and 1.5), are internal deliverables that follow the progress of the OCR work of the 14 content providing partners of WP1.

The other two documents, the "Survey of Availability of Digitised Images for OCR" and the "Survey of Existing OCR Practices and Recommendations for More Efficient Work" (TELplus deliverables D 1.1 and D 1.2), are public deliverables whose purpose it is to provide access to the findings of this workpackage to the wider public as well as to the TELplus partners.

In January 2008, the results of the "Survey of Availability of Digitised Images for OCR"² were published. For this and for the "Survey of Existing OCR Practices", a combined questionnaire was sent out to the 14 content partners in TELplus WP 1. The first survey was aimed at the collection of temporal, geographical and linguistic information about the original documents OCR'ed for TELplus. Some information in that first survey was already related to the OCR workflow at the respective national libraries.

The "Survey of Existing OCR Practices" will focus on more practical aspects of OCRing (software used, OCR accuracy) and of giving access to OCR'ed fulltext through libraries' websites (search engines used, integration into Digital Object Management (DOM) systems). Some information collected for D 1.2 was needed earlier in the project to plan for task 1.5 "Provision of Access to Newly OCR-ed Material through The European Library". For this reason, it was decided to release that information already with the first survey. It has been updated, where necessary, and is published again here. This is the only way to achieve a good overall picture of the existing OCR approaches in the partner libraries.

The current document will be divided into two parts. The first will focus on the results of the survey, while the second will draw on these results and on other input to provide recommendations for dealing with common challenges in OCRing digitised library material.

The survey part of this document is aimed mainly at getting a better picture of the TELplus WP 1 content providing partners' OCR work and the way they will provide access to their material from their own website. The recommendations, while in the first instance aimed at furthering the OCR work inside TELplus, are designed to help anyone planning an OCR project in considering the important points of such a project.

² See [13]

1.2 Input for this document

As has been said above, much of the information for the survey part of this document was gathered with a combined questionnaire for both surveys for TELplus workpackage "Making Searchable Digitised Images via OCR" (WP1).

Additional information was gained through a template the partners were asked to fill in for D1.3 and through personal contact between the workpackage leader and individual partners. From this, a fuller picture could be drawn of the partners' approaches to OCRing and to making their OCRed material accessible through their websites.

One important source of information for the good practice recommendations was the input from a workshop on OCR practice organised by the Bibliothèque nationale de France and held on 28 and 29 January 2008 in Paris. Here representatives from almost all content providing partners of WP 1 gathered to exchange experience, listen to, and discuss presentations on typical OCR challenges.

Further input came from colleagues from other European national libraries, namely the Koninklijke Bibliotheek of the Netherlands, a representative of which was also an invited speaker in Paris, and the British Library.

A final source of information was desktop research, gathering information about OCR projects, their challenges, and how these have been coped with.

2 Survey of existing OCR approaches

Many libraries have digitised content that is not searchable in detail. This material usually consists of digital images of original pages of books or other publications that are provided online with little or no descriptive metadata for the lower levels of the document structure. There is no possibility to search in the text of those documents.

The goal of TELplus WP 1, “Making Searchable Digitised Images via OCR”, is to use digitised images of historical texts for the creation of fulltexts via OCR. The fulltext produced will be used for detailed searching in existing digital libraries; thus, a considerable added value will enrich the existing data.

This fulltext material will be delivered to The European Library by 14 European national libraries³, each with their own background and experience. To get an overview of the different approaches to OCR the partner libraries have taken a survey was conducted.

The price each library calculated for their OCR has been collected, even though this is not directly part of the OCR approach of the library. However, many discussions about OCR workflows turn at some point to the price question. Sadly, the sample produced here is too small as to give a comprehensive picture of the situation.

The chapters on OCR software, output format and quality assurance are the ones that deal with the actual workflow.

Further chapters deal with the individual setup the TELplus partners use to give access to their fulltext material. This, too, is not really a part of the OCR approach, but the way the library will be able to display their fulltext material is dependent on decisions made in planning the OCR workflow. Additionally, this information is important in future planning for task 1.5 "Provision of Access to Newly OCR-ed Material through The European Library".

³ See table below.

Table of TELplus WP1 content-providing partners

National Library	Short name	Country
Austrian National Library	ÖNB	Austria
National Library of the Czech Republic	NLP	Czech Republic
National Library of Estonia	RR	Estonia
French National Library	BnF	France
National Széchényi Library of Hungary	OSZK	Hungary
National and University Library of Iceland	LBS-HBS	Iceland
National Library of Latvia	LNB	Latvia
Martynas Mažvydas National Library of Lithuania	LNLM	Lithuania
National Library of Norway	NLN	Norway
National Library of Poland	BN	Poland
Slovak National Library	SNK	Slovak Republic
National and University Library	NUK	Slovenia
National Library of Spain	BNE	Spain
National Library of Sweden	NLS	Sweden

2.1 OCR pricing

To get an idea of the price range paid for OCR across the partners, prices for a single OCRed page are collected here. To make these prices more comparable, and because most libraries currently digitise and OCR newspapers⁴, we specified the price for an OCRed newspaper page for this list.

Those partners who are doing their work in-house were asked to count only the cost of licences and personnel needed for the workflow of OCR and not that of administration, file

⁴ See also [13]

management and so on. These costs have to be taken into account for all OCR work, regardless of whether it is done in-house or via a contractor. An important cost factor for in-house OCR will always be the hardware. This, however, is also the most difficult factor to break down for a single page. For this reason, it was decided not to include hardware in the in-house page price.

Not all partners were able to give a price for the OCRing of a single newspaper page. To get a more rounded picture, the responsible colleagues from the British Library (BL) and from the Library of The Netherlands (KB) were also asked to give the price they pay per page.

Table of OCR page prices calculated by TELplus WP1 content-providing partners

National Library	Out sourced/In-house	Price per Page
ONB	out sourced	0.25 €
NLP	out sourced	0.026 € - 0.032 €
BNE	out sourced	0.04 € - 0.06 €
BnF	out sourced	0.063 €
KB	out sourced	0.15 €
BL	out sourced	0.32 €
LNLM	out sourced	0.61 €
OSZK	in-house	0.05 € - 0.10 €
SNK	in-house	0.10 € - 0.50 €
RR	in-house	0.20 € - 0.66 €
NUK	in-house	0.30 € - 0.65 €

The table shows, that in such a small sample of libraries, the prices vary enormously. This has a number of reasons. Firstly, prices for contracted work dependent on the amount of pages to be OCRed within a single contract. Similarly, the probability of future contracts is relevant. Secondly, the quality, size and complexity of the digitised originals and the type of output a library requires will influence the price they pay for OCR.

The National Library of The Netherlands, for example, gets their output in three files. They pay 0.07 € per page for a plain OCR xml file (with an average of 14 articles per file). Another 0.07 € are paid for an ALTO file per page, and an additional 0.14 € for a PDF per issue (with an average of 12 pages per PDF). Thus, the price per page for all required files comes to about 0.15 €.

The National Library of the Czech Republic, on the other hand, will receive simple text (provided as PDF) and additional METS files. The price range for their OCR is due differences in the size of the original material. The library pays 0.026 EUR for pages up to A2 size and 0.032 EUR for pages of A2 size and bigger.

The highest price in the group that contracts their OCR work is being paid by the National Library of Lithuania. The Lithuanian newspapers were digitised from bad quality microfilms, resulting in images that are also of poor quality. In order to achieve high quality OCR results all recognised pages will be revised manually by the subcontractor. This in turn results in an exceptionally high price, but it will also make it possible to provide the users with readable text.

For in-house work, a similar price range is apparent. If anything, the top prices are slightly higher. It has to be kept in mind that they are estimated prices and quite large cost factors have been left out of the reckoning. This makes these prices rather unreliable, whereas the prices for contracted OCR work are usually fixed prices.

So from this small sample it seems that the decision whether to OCR digitised material in-house or to contract it out does not matter too much when it comes to the price of OCR. One must bear in mind however, that in-house work is very much dependent on the soft- and hardware that is needed to OCR digitised text. The basic initial investment the library must make will only pay off if a sufficient amount of data is produced.

2.2 OCR software

The range of OCR software fitted to the needs of libraries is rather limited. In fact, all TELplus fulltext is produced using some version of Abbyy's FineReader (<http://finereader.abbyy.com/>). Besides versions 5.0 to 8.0, versions XIX (for Gothic (Fraktur) fonts) and the Recognition Server are in use.

Those libraries or subcontracted service providers that do not use Abbyy's products directly have used some sort of derivative. The documents of the National Library of Norway and the Austrian National Library are OCRed using docWORKS, a CCS product (<http://www.ccs-gmbh.de/de/digitization.htm>). docWorks in turn is based on a variety of FineReader versions. Similarly, the former service provider of the French National Library used its own product, which was in turn based on a combination of OCR products – FineReader among them.

Accordingly, the results of fulltext production are comparable and depend more on the choice of original material than on the quality of the software. The Martynas Mažvydas National Library of Lithuania and the National Library of Poland get almost perfect results, because both do manual revision after OCR.

Table of OCR software in use by TELplus WP1 content-providing partners

National Library	OCR - software
ONB	docWorks
NLP	FineReader 5.0 , 6.0, 8.0
RR	FR 8.0, XIX, Abbyy Recognition Server
BnF	FineReader 8.0
OSZK	FineReader 6.0-8.0 (9.0)
LBS-HBS	FineReader 7.0
LNB	FineReader 8.0
LNM	FineReader 8.0 + manual revision
NLN	docWorks
BN	FineReader 8.0
SNK	FineReader 0.7-0.8
NUK	FineReader 8.0 pro / FineReader development versions
BNE	FineReader 8.0
NLS	FineReader 7.0

2.3 OCR output formats

After choosing the material for OCR and deciding on the method used to OCR it, the decision on the output format⁵ is the next important step. There are a number of considerations regarding the display of the document, but also about the size of storage needed, involved in this decision.

As only two of the libraries do manual revision of their OCR results, most of the others have decided not to display their fulltext directly to the user. In fact, only the national libraries of the Czech Republic and of Poland do so as their first choice of access to their fulltext. Three libraries give their fulltext via HTML as an alternative form of access. Of these, only the National Library of Poland does manual revision.

Three methods let the user see the actual image of the original page. In a METS/ALTO setting, with DjVU or with PDF the fulltext is usually only used to search the text and to

⁵ This chapter is written with the assumption that the reader has some idea of typical OCR output formats. For a description of these formats, please refer to the corresponding chapter in the Good Practice Recommendation part of this document.

display the results of the search on the digitised page (highlighting). However, unless appropriate measures are taken, the user will normally find a way to access the fulltext⁶.

Out of the 14 partners that provide fulltext material to TELplus, the National Library of the Czech Republic and the National Library of Latvia are using DjVU as an alternative access format. The National and University Library of Iceland, who has until recently given access to all their OCRed material in the DjVU format is currently changing to PDF. One main reason for this is that most modern computers already have a viewer for this format installed, while a viewer for DjVU will in most cases have to be installed to view the material.

Nine of the 14 partners give access to their material in PDF. While this is not the main access format in all libraries, it is at least the one, which has the widest use within this group. This is especially true as two more libraries are considering PDF as an alternative for – at least part of – their material. Additionally, the French National Library produces PDF from its METS/ALTO files in a second step. In this group, the Martynas Mazvydas National Library of Lithuania is special in that they give access to the digitised images as PNG or JPEG, while access to the fulltext is given as PDF.

To display the image with an underlying METS/ALTO structure is probably the most elaborate way to give access to the TELplus material. Five partners use METS/ALTO setups as the main way to give access to their material. Of these, two (with the French National Library, three) use PDF as an alternative form of giving access to their fulltext. This lets the user take away the documents, which is not possible with the METS/ALTO structure.

⁶ For more detail, please see discussion below.

Table of OCR output formats used by TELplus WP1 content-providing partners

National Library	OCR output format(s) used
ONB	METS/ALTO (PDF for newspapers - not yet decided)
NLP	DjVU (potentially also PDF)
RR	PDF
BnF	METS/ALTO (PDF produced from METS/ALTO and Image)
OSZK	PDF
LBS-HBS	Changing from DjVU to PDF with all new material provided in PDF
LNB	PDF, HTML
LNМ	PDF
NLN	METS/ALTO
BN	Text (BN is considering PDF and HTML as possible future output)
SNK	PDF
NUK	METS/ALTO, PDF, HTML
BNE	PDF
NLS	PDF and HTML

2.4 OCR quality assurance

After the text of a digitised document has been recognised, most of the partner libraries do some kind of quality assurance⁷. Almost half of the partners simply use the software log, a file in which the OCR software documents whether a letter has been recognised correctly according to the software's algorithm.

Most of the other partners do human-eye spot testing, in which a random sample is checked by comparing the text digital image to the fulltext. Both methods have their specific challenges and neither gives very accurate results.

Only the National Library of Poland and the Martynas Mažvydas National Library of Lithuania estimate their results as 99% accurate, but this is because their OCR results are revised manually.

⁷ This chapter is written with the assumption that the reader has an idea of possible methods OCR quality assurance. For a description of these formats, please refer to the corresponding chapter in the Good Practice Recommendation part of this document.

Another factor that influences the established rate of accuracy of OCR is the error counting method. Generally, it is possible to count either correct vs. incorrect letters or correct vs. incorrect words. Most partners have decided to count letters.

Table of OCR quality assurance used by TELplus WP1 content-providing partners

National Library	OCR output format(s) used
ONB	human eye spot test/letter count
NLP	human eye spot test/letter count
RR	software log/letter count
BnF	human eye spot test/letter count
OSZK	human eye spot test/word count
LBS-HBS	human eye spot test/letter count / human eye spot test/word count
LNB	undecided
LNM	manual revision
NLN	software log/letter count
BN	manual revision / human eye spot test/letter count
SNK	software log/letter count
NUK	software log/word count / none
BNE	none
NLS	software log/letter count

2.5 Digital Library Systems

One of the topics that does not directly relate to an OCR issue but which are interesting in the context of TELplus is that of the Digital Library Systems or Digital Object Management (DOM) systems used by the content-providing partners. Part of this topic is the question about the search solution used to provide fulltext search.

From the table below, it can be seen that the TELplus partner libraries use a large variety of DOM systems. In fact, more or less every library has its own Digital Library System. Only the BNE and the ONB (decision pending) have DigiTool as one option. Four of the libraries use systems based on the Fedora Open-source Digital Object Repository Management System.

Most of the libraries in TELplus WP1 use systems that were built either by library staff or specifically implemented for the library.

This diversity, much more than the variety in OCR output, makes detailed planning for WP 1 task 1.5 "Provision of Access to Newly OCR-ed Material through The European Library" important. This task is designed to make sure that all material OCR-ed for TELplus WP1 will in the future be accessible through The European Library. Additionally, task 1.5 is working on a proof-of-concept for an index over the total of more than 2 million pages that will be provided by TELplus WP1.

Table of DOM systems used by TELplus WP1 content-providing partners

National Library	Digital Object Management system used
ONB	DigiTool (http://exlibrisgroup.com/digitool.html) or Greenstone (http://www.greenstone.org/), not yet decided
NLP	System build for the library, now used by a number of Czech libraries (http://www.qbizm-technologies.cz/Reference/pripadove_studie/narodni_knihovna/index.html)
RR	Fedora (http://fedora-commons.org/)
BnF	SPAR , a system designed in-house, based on Fedora (http://www.bnf.fr/pages/version_anglaise/numerisation/num_spar_eng.htm)
OSZK	Multiple systems build in-house. A new system OSZKDK is currently being set up and is supposed to supersede the older systems.
LBS-HBS	In-house solution based on a relational database
LNB	Solution by Olive Systems (http://www.olivesoftware.com/)
LNM	LIBIS (Lithuanian Integrated Library Information System) (http://www.libis.lt/) based on the Oracle DB 10g database (http://www.oracle.com/database/index.html)
NLN	System build in-house
BN	dLibra (http://dlibra.psnc.pl/index.php?lang=en), a Polish library solution
SNK	VITAL (http://www.vtls.com/products/vital/), a Slovak library solution based on Fedora
NUK	dLib, a system build in-house
BNE	DigiTool (http://exlibrisgroup.com/digitool.html) and PANDAS (http://pandora.nla.gov.au/pandas.html)
NLS	Fedora (http://fedora-commons.org/)

2.6 Search solutions

Directly connected to the question of which DOM system the libraries use is the question of the search solutions libraries employ to search their own fulltext collection.

The use of search engines used to index and search the fulltext collections is much more homogeneous than that of DOM. Out of the 14 partners in WP1, nine are either using or are in the process of implementing a search solution based on the open source search engine Lucene.

Most of the other libraries use either a search solution build into their DOM or into the database that underlies it.

Table of search solutions used by TELplus WP1 content-providing partners

National Library	Search solution used
ONB	Lucene (http://lucene.apache.org/)
NLP	Lucene (http://lucene.apache.org/) and Convera (http://www.convera.com/)
RR	Lucene (http://lucene.apache.org/)
BnF	Lucene (http://lucene.apache.org/)
OSZK	Search++ (http://swishplusplus.sourceforge.net/) (Lucene for the new system OSZKDK)
LBS-HBS	Lucene (http://lucene.apache.org/)
LNB	Build-in proprietary solution
LNM	Oracle Text (http://oracle.com/technology/products/text/)
NLN	FAST search engine (http://fastsearch.com/)
BN	Not yet decided, probably Lucene (http://lucene.apache.org/)
SNK	Build-in proprietary solution
NUK	MS SQL FTI (FullText Indexing) (switching to Lucene)
BNE	Build-in proprietary solutions
NLS	Lucene (http://lucene.apache.org/)

2.7 Access points

Most of the partner libraries have special access points for their digital collections. Probably the best-known example of this is the Gallica website of the French National Library. Some of the partners give special access to their digitised text material, like the Czech Kramerius or the Icelandic Timarit websites. Only three partners give access to their digitised material through separate access points for each type of material. The Austrian National Library, for example provides access to its digitised newspapers via the ANNO and to its governmental

material via the ALEX sites. The National Széchényi Library of Hungary and the National Library of Spain even use different DOM systems for each type of material. However, both libraries are now in the process of unifying their systems and will give access to all of their material through their main collection sites in the future: The Biblioteca Digital Hispanica is the Spanish digital library and will soon also display their newspapers, which are currently accessible via the Hemeroteca Digital. The Hungarian digital library Digitális Könyvtár will integrate both the books and the periodicals, which currently run on two older systems.

Table of links to TELplus WP1 content at partners' website (search access)

National Library	Digital Object Management system used
ONB	http://anno.onb.ac.at for newspapers and http://alex.onb.ac.at for governmental material
NLP	http://kramerius.nkp.cz
RR	http://digar.nlib.ee/otsing/avaleht?stype=c
BnF	http://gallica2.bnf.fr/
OSZK	http://www.elib.hu for books and http://efolyoirat.oszk.hu/ for periodicals. All other material can be searched from http://oszkdk.oszk.hu/search/
LBS-HBS	http://timarit.is
LNB	http://periodika.lndb.lv or http://www.periodika.lv
LNLM	www.epaveldas.lt/vbspi/content/simpleSearch.jsp
NLN	http://www.nb.no/sok/search.jsf?query=
BN	http://polona.pl
SNK	http://www.memoria.sk
NUK	http://www.dlib.si
BNE	http://bibliotecadigitalhispanica.bne.es/R/ and http://hemerotecadigital.bne.es/ for newspapers
NLS	http://www.kb.se/

2.8 Summary

This survey is the second of a set of two surveys that were planned as input for TELplus WP1. Its purpose was to give an overview of the different approaches to OCR the partner libraries have taken. Together with the "Survey of Availability of Digitised Images for OCR", this survey is designed to give a good general idea of the work done in this workpackage.

The surveys inform about the linguistic, geographic, and temporal scope of the material that is being OCRed for TELplus, about the different workflows that have been established at the partner libraries, and about the way, the material is accessible from the libraries' websites.

This input has been essential in providing basic information in planning the workpackage. Much of the information collected for these surveys has been and will in the future be important in planning task 1.5 "Provision of Access to Newly OCR-ed Material through The European Library".

Additionally, many of the suggestions given in the good practice recommendations below have their origin in challenges the partners have faced and the solutions they have found to cope with them.

Finally, both surveys will give interested outsiders an overview of the project and of the participating partners.

3 Good practice recommendations

Digitised images of textual (and other) documents are a way to give access to original documents to people who cannot (e.g. because they are not at the library) or should not (e.g. because the documents in question are fragile or extremely valuable) handle the originals. In the way they are used, these digital copies are in most cases no different from the originals: Users have to have an idea of which documents they need to search for the information they need. To find the parts they are looking for, they will have to leaf through the documents (and use tables of content, chapter headings, and indices). Optical character recognition (OCR) is a means to enhance access to these documents. With the additional provision of a document's fulltext, it becomes searchable. Documents that are not an obvious choice for a user looking for e.g. a certain phrase, location or name, will still be found by the search engine. This gives the user access to seemingly unlikely documents as well as pointing them directly to the place in the document that contains the words they were looking for.

Today, supplying OCR results as part of a digitised library document (be it from a newspaper, book or manuscript etc.) is usually a question of access (as described above) rather than of preservation. This may change in the future when the quality of character recognition, image enhancement, text correction, and other factors important to the quality of OCR results have progressed to a stage where 100% correct fulltext can be expected as a typical part of the digitisation of older documents. Currently, however, most libraries only supply OCR results as a means to enhance access to their digital material. Some libraries will not even show the OCR results to their users, others post notices on their web sites, warning their users not to expect too much from the added fulltext. This is the light in which the following good practice recommendations for the OCRing of older text material have to be read.

As digitisation and OCR are closely linked, there are only few documents currently to be found that deal solely with challenges arising in OCR workflows. Most will consider either only digitisation or OCR in conjunction with the digitisation workflow. To some degree this last will be true also for this document. While the focus of the recommendations is the OCR workflow, many decisions that affect the OCR work will have been made while digitisation of the originals was planned. While some recommendations that are meant to be considered when making these earlier decisions have been added here, others have been left out. This is especially true for recommendations about copyright issues, which are important to keep in mind⁸; they have been dealt with extensively by others.⁹

3.1 Project purpose

It is important to notice, especially in the light of big digitisation projects like those by Google, Microsoft and others, that mass digitisation is not the only possible purpose of a

⁸ A number of partners in the TELplus project have had IPR problems. This shows that it is very important to check all material for IPR, but this should already be done when the choice on material for digitisation is made.

⁹ See for example [12]

digitisation project. This is especially so, because the purpose of such a project will inform many of the choices taken in its preparation, but it will also be expressed by the projects size. Besides large scale and mass digitisation, which are not necessarily the same, preservation is a common reason to start a digitisation project. These projects usually involve a lot of preparation because the documents chosen for such a project are either deteriorating or of special value, meaning that they will have to be handled with great care. Documents of such projects may also be OCRed, but they will often not yield very good results for reasons laid out below. However, even comparatively low OCR quality may broaden access to such documents, because even these documents will be found in some searches. This is important, because the original chosen for such projects are often the most used in a library, which may be one reason for their deterioration.

Some libraries digitise special collections. Here, originals that either are deemed important in a particular setting or are valuable may be chosen. In these projects, special care is often put on the quality of the digital copy. Scan will be prepared with regard to the original, but also to its faithful depiction. This means that such scans may lend themselves to OCR, especially if there is low font variation and the OCR software is specially trained.

Both preservation and special collection projects have a tendency to be smaller projects in which more care is taken in selection and preparation of the originals.

Large-scale projects will usually focus on the amount of digitised material that is produced. Here the choice and preparation of the originals will be less important. Among the larger project mass digitisation projects, which often intend to digitise "all" of a certain set of originals, will place the least amount of effort in the quality of the digitised material. Here the final amount is the main overall goal.

These categories can of course, only give an idea of the differences in the purpose of digitisation projects. Often more than one goal is important, so that a project in which "all" of a library's endangered and often used material is being targeted may not be small and may place relatively little effort into the actual choice of documents.

3.2 Project size

Of course, the size of the project will matter when planning an OCR workflow. Large-scale projects, in which millions of pages are OCRed, will often have a completely different regard for some factors of the OCR workflow than smaller projects. Because of this difference, recommendations for each type of project (large scale or small, rather than by purpose) are given separately where this seemed sensible.

When it comes to the size, OCR projects at European national libraries range from projects that more or less "digitise and OCR everything" in text based material the library possesses to relatively small projects that OCR 100.000 pages at one go.

In Hungary, for example a list of important Hungarian literature has been established along with the guidelines for digitisation. Before digitisation, a systematic comparison is done to see whether a different library already offers a digital copy of a specific document. Periodicals are

only digitised if the collection is in a good physical state and no microfilm is available for the title.¹⁰

In Iceland and Norway, on the other hand, there are programmes to digitise the whole of the libraries' holdings and most of the material is being OCRed. In Norway this programme is not only focused on text material. Here all analogue material is being selected for digitisation. This includes audio and visual (e.g. radio and TV) material as well as all printed material like newspapers, books or maps.

3.3 Choice of material for OCR

Regardless of the size of the project in question, some general points will have to be considered. First among them is the question of how the fulltext will be displayed.

The simplest use would be an index of all created fulltext. This makes it possible to add search functionality to the collection and to direct users to the documents or parts of documents in which their search terms can be found. For such an index to yield usable search results, the OCR result can be of relatively poor quality. This has two reasons: Firstly, the user will often never actually see the fulltext and will thus have no reason to complain about too many mistakes. Secondly, the search, which in itself is already an added value to the digitised text, will often still work, because important words usually appear more often in the text. Therefore, even if one instance of the word is not recognised correctly a second instance may be. Thus, the users may still find their words in the text. Of course, the better the text was recognised, the better will the search results also be.

Accordingly, for this first type of OCR use, the choice of material may well be very general.

The most demanding use of OCR results is the display of the fulltext either as the only method of displaying the resource or as one alternative of displaying it. The advantage from the user's perspective would of course be that they can copy and paste the text and that they will find any word they are looking for, anywhere in the text (even when the search engine cannot find it). For this, the library will need to make sure that the recognised text has as few recognition errors as possible. Displaying the text in the quality that is currently produced will not only make it difficult to read, but may lead to a loss of trust in the reliability of the resource and hence of the library that offers it. There is currently no way of producing this level of correctness in fulltext from a digitised document using only OCR software. This means that a library that wants to display their fulltext will have to consider either keying (re-typing) the text or manually revising the OCR results. Both methods are of course infinitely more expensive than simply using OCR software. There are already some methods of post-correction (see the chapter on post-processing), but these will currently not produce the same results as either keying or manual revision.

Most libraries will use their fulltext in a setting that is somewhere between the two described above. They will often use OCR to add searchability and missing structural metadata to their

¹⁰ This is, of course, a question of the purpose of the project, but such intensive pre-selection is only really possible for comparatively small projects.

digitised collection. By using certain output formats, especially PDF, users will still have access to the fulltext, even though it is not given them directly. A simple "save as text" will extract the fulltext from the PDF file, unless the file has been secured to prohibit this. Even then a more skilled user will still be able to get the text. (See also the section "Choice of OCR output format", below)

One possible way of giving access to fulltext as an alternative to display the digitised image without losing the users' trust may be to put up a warning notice like the following:

FAQ section of the newly released Times Archive:

Why does some of the OCR text look messy?

The quality of text that is extracted through OCR depends on the quality of the image and on how easy it is for the process to recognize the characters. Some of the older newspaper pages have become a bit battered over the years and will produce lower-quality results, as will older fonts. In the early years of The Times the character s was commonly printed as an f, which will confuse the spelling in the OCR version.¹¹

Technology section of the Papers Past website of the New Zealand Digital Library Project:

OCR mistakes

OCR enables searching of large quantities of fulltext data, but it is never 100% accurate. For the Papers Past project, the level of accuracy depends on the print quality of the original newspaper, its condition at the time of microfilming, and the level of detail captured by the microfilm scanner. Newspapers with poor quality paper, small print, mixed fonts, multiple column layouts, or damaged pages may have poor OCR accuracy. This means that most pages will have some errors in the computer-generated text, and some will have a lot of errors.¹²

Both examples are not placed very prominently on the respective website. The FAQ section is possibly a somewhat more likely place for a normal user to look for such a notice than the technology page. It may a good idea though to put such a notice, or at least an obvious link to it, into a more visible spot on the website. Thus it may act as more of a warning to use this text as it is found, but not to trust in it as a source of reference.

3.3.1 Large scale projects

Oya Rieger says¹³ that "[t]he primary aim of large-scale digitization projects—to quickly create a critical mass of digitized books—stands in contrast to that of earlier projects, which frequently sought to create fewer, but higher-quality, scans for scholarly use." To achieve this

¹¹ [23]

¹² [15]

¹³ [17] page vi.

critical mass, she says, it is not advisable to put too much effort into the choice of material. Choosing material for digitisation and OCR is labour and hence cost intensive and money thus spent may be better invested in more pages of digitised and OCRed material, rather than in devising intricate criteria for choosing material. If a library is not in a position to digitise all or the majority of its textual collection, choosing certain sub-set of that collection may be the most efficient way. This could be according to period (e.g. any thing between 1850 and 1900) or type of material (newspapers, scientific papers, PHD theses etc.).

The best idea, especially for libraries changing from the more traditional approach to digitisation to that of mass-digitisation¹⁴, may be to start with books¹⁵. Books are usually the least problematic material as regards segmentation. They are normally printed in a limited set of fonts, and offer, apart from indices and chapter headings little difficulty for OCR software.

Many large-scale digitisation projects today focus on newspapers. For this reason, a lot of work has been put into advancing the segmentation abilities¹⁶ of current OCR software. For this reason newspapers may also be a good choice for a mass digitisation and OCR programme. As opposed to books, though, some work will have to be put into adapting segmentation templates that are used. Also, the results of segmentation has to be more closely monitored than for books; and a much larger variety of fonts are found in newspapers than in normal books.

3.3.2 Smaller projects

In many smaller projects, the value or importance of the material to be digitised (and OCRed) is given as the most important aspect of choice for small-scale digitisation projects. Usually these are actually documents that are most used or most prone to decay. In libraries where the collection is well known and documented, a well thought up system for choosing material may be very important for the overall result. It should however be remembered that, as in large-scale projects, the effort and money spent on the selection process should be limited to the necessary. This way more documents can be digitised and OCRed at the same time and with the same amount of money.

In the end, some decision on which material to OCR will have to be taken. There are some points, which may generally be considered in such a decision. As has been said above, a general decision may be on the type of material to be OCRed. Here, as with the same decision in larger projects, the easier the material the better the quality that is to be expected. Besides books and simple printed manuscripts, newspapers present themselves as a good choice. They are the most common type of document in such projects and much research has gone into dealing with the challenges of this kind of material.

If possible, the choice should be made in such a way that variations in layout, font and language in the material changes as little as possible as each of these will call for adjustments in the OCR software¹⁷.

¹⁴ That is, one where the amount of digitised pages counts.

¹⁵ Cf. [25]

¹⁶ See the chapter on pre-processing below.

¹⁷ For more detailed information on this see e.g. [07] and [21].

Most important is, of course, the quality of the original material. Dirty or torn pages as well as imperfect printing are problematic for the OCR software. However, these may vary over a given set of documents, and checking and potentially discarding a number of such sets may take a lot of effort.

3.3.3 OCR input

Currently most OCR-software vendors will ask for greyscale images as input for their programme to work well. Very good bi-tonal images will often work as well.¹⁸ Binaryisation¹⁹ is usually a part of the work the OCR software does as pre-processing, but the problem here is that the human eye works differently from the software. This means that what looks like a good bi-tonal source image may not yield good quality OCR results.

Regarding the resolution of the images, it seems to be a generally accepted that 300 dpi is a good quality for human readability as well as for OCR. A higher resolution will most likely increase the recognition quality on images from a bad source (e.g. dirty pages or small fonts), but will also increase the size of the digital images.

3.3.4 Digitised microfilm material as input for OCR

Many libraries have large stocks of microfilmed text material and some have started to use these films as source material for digitisation, rather than the paper originals. This has the advantage that such digitisation can be done by scanning machines, saving working time and money.

The problem with digitised images from older microfilms is that they are usually not suitable for OCR. There are a number of reasons for this. The main reason is that each copy one makes of an original is "lossy", that is, there are mistakes in the copy. A copy from a copy will of course be even lossier than one from an original, as each method of copying has its own type of loss. Thus every new copy will accumulate the old loss and the new.

A second reason why digitising old microfilms is not a good choice for OCR is that microfilms deteriorate over time. The result, again, is additional loss. What is more, old microfilms were never intended to be machine-readable. This means that skewed or slightly blurry images were not a big problem as people could still read or print them. For OCR, however, skewed or blurry images are major problems.

There are, however libraries, like the National Library of the Czech Republic, which first produce microfilm and then digitise from this. These microfilms have, however already been made with subsequent digitisation and OCR in mind and may be better suited than older microfilms.

If the decision to use microfilm is made, it is important to start with a test run to evaluate the possibility of OCRing that material.

¹⁸ Grayscale images are black-and-white images as they are known from black-and-white photography, meaning that they actually display gray tones. Bi-tonal images, on the other hand, are really black and white as they display only these two colours and no grey tones.

¹⁹ See the chapter on pre-processing below.

3.3.5 Avoiding duplication of work?

There have been discussions about the use and the possibility of building databases of digitised (and OCRed) texts.²⁰ While such databases may make sense in the American or even the Anglo-American context, they certainly are not yet sensible in the context of European national libraries. Many of these libraries have not yet begun to digitise large amounts of material and the focus they have in choosing their material is always a national one. While there may be a common interest in the selection of material between countries sharing borders and/or parts of their histories, it seems, at least for now, much more logical to work together in a bi- or even a multi-lateral setting rather than via a large database. Especially as such a database would slowly have to grow in the future.

Checking for already digitised OCRed material in other libraries for the purpose of choosing what to OCR or for avoiding the duplication of work is currently too time consuming. It may well be a better use of a libraries funds to digitise and OCR those parts of its assets it chooses and risk the small amount of duplication that may occur.

3.4 Pre-processing

There are a number of possible steps that can be taken to pre-process the digital images in order to improve their readability for OCR software. These are often built into the OCR software, which is good, because all changes should be made in accordance with the needs of that specific software. However, any changes the user makes in such combined software happen in a kind of black box that gives no direct insight into what the changes actually do.

Of the changes that are possible, the following are probably the most usual ones²¹:

- Binarisation:** Binarisation is the conversion of a greyscale (or colour) image into a black and white or binary image. OCR software will usually binarise them prior to the actual OCRing. The resulting bi-tonal image will not be produced for human readability, but will rather support the software's work.
Binarisation may help reduce the effects of e.g. letters shining through, of bleed through. Some of the following methods may only work properly on bi-tonal images.
- Straightening:** Straightening corrects the angle of a page that has been scanned at an angle to the scanner's sides.
- Dewarping:** Warping happens during the scanning process. It produces curved lines. Dewarping can straighten such lines, increasing both human readability and OCR recognition.

²⁰ Cf. Carney's, Schwarz' and Stockmann's presentations given at at the LIBER-EBLIDA Workshop on Digitisation at the Royal Library, Copenhagen, Denmark 24 October 2007. [03] [18] [20]

²¹ The various methods of pre-processing explained here are currently research topics in the IMPACT project (<http://www.impact-project.eu/>). IMPACT (IMproving Access to Text) is funded under the Seventh Framework Programme of the European Commission (FP7) (<http://cordis.europa.eu/fp7/ict>)

Removal of boarders: Around many digitised images there are black borders, around either the whole or part of the image. Their removal is not only a "cosmetic" operation, but may enhance OCR accuracy as such boarders have a tendency to be recognised as characters.

Page segmentation: Page segmentation divides the page into headers and text, individual paragraphs, columns, and, e.g. in newspapers, into individual articles.

Though there are individual tools for some of these tasks, there have been few large-scale tests for those tools. This means that most current OCR projects rely on the abilities of the OCR software. It is therefore advisable to check whether the software to be used is able to execute any or all of these methods.

3.5 Choice of OCR output format

Basically what OCR produces is text. This may come in simple text files, which are either used only for searching that OCR'd text or they may be presented to users so that they can access the fulltext directly. The text may also have mark-up added to it, as in an XML file. Here the user will be presented with text either exclusively or as an alternative to the digitised image of the original.

Two more elaborate possibilities to output OCR results are PDF and METS/ALTO. In both cases, the user will be presented with digital images to which the recognised text is connected in such a way that searching and highlighting of the search results on the digitised image is possible.

In a PDF,²² this is achieved by including digitised images and the corresponding text in one file and adding information about the position of the individual words on the images. These files are easy to download and transport, but, depending on the amount of pages and the quality and size of the images they contain, have the tendency to be very large. It has to be kept in mind that the PDF contains the OCR results and thus the user usually has access to it. This means that the issues concerning the display of the fulltext that have already been discussed will also have to be considered with PDF.

METS/ALTO is a combination of two types of metadata files that was originally designed for displaying digitised and OCR-ed newspapers. The METS²³ file contains metadata about the physical and/or structural layout of the digitised original. This may in the simplest case be the order of the pages of a digitised publication, but may also give more detail. It may, for example, identify individual articles in the publication or even sections, like weather sections, obituaries and so on. The METS file also documents the relationship between digital files. In this case the relationship between the digital images and the corresponding ALTO files. The ALTO²⁴ file, on the other hand, contains the OCR-ed text and structural information about the individual pages of the digitised publication. In it may be found the position of headlines, paragraphs, lines, and especially of individual words on the page. The combination of METS

²² Portable Document Format see: http://www.adobe.com/devnet/pdf/pdf_reference.html

²³ Metadata Encoding and Transmission Standard see: <http://www.loc.gov/standards/mets/>

²⁴ Analyzed Layout and Text Object see: <http://www.ccs-gmbh.com/alto>

and ALTO makes it possible to not only search for and highlight individual words on the page, but also whole sections or articles, even parts of articles, images, tables, or charts, all according to the setup used.

Of course the combination of METS and ALTO files plus the corresponding images is not easily downloadable and portable because they are designed for display in a browser. The library may additionally provide PDFs (or simple text files) and offer them for download. With the right setup PDFs (or again simple text files) may also be produced on the fly. These PDFs may be much smaller than those created directly from the OCR-ed data, as the setup may allow for the creation of PDFs of individual pages, of articles or even collections of articles. Again, this has to be setup and made possible by the library beforehand.

An output format that was originally designed especially for digitised and OCR-ed data is DjVU (pronounced *déjà vu*)²⁵. As with PDF a special program (or browser plug-in) is needed to view DjVU files. As both PDF and DjVU work in similar ways and as most users already have a PDF viewer installed, DjVU is currently not a preferred output format for most libraries. In fact, some libraries are currently changing their output from DjVU to PDF to make access easier for their users.

The majority of the libraries in the workpackage 1 of TELplus will produce PDFs as their output. Some libraries have decided produce more than one kind of output for each original document.

3.6 Measuring OCR output correctness

Once the OCR results have been delivered, the library will want to get an idea of the quality of the recognised fulltext. There are several way of doing this and a number of considerations to be taken²⁶.

The quality of OCR results can be checked in a number of different ways. The most effective but also most labour extensive method is manual revision. Here the staff of the library or the subcontractor checks the complete OCR result against the original and/or the digitised image. While this is currently the only method of checking the whole OCR-ed text, and the only way to get it almost 100% correct, it is also cost prohibitive. For this reason, most libraries reject it as impractical.

All other methods of checking the correctness of OCR output can only be estimations, and none of these methods actually provides better OCR results. That is, further steps, which will include manual labour, will have to be taken to receive better results.

3.6.1 Software log analysis vs. human eye spot test

To get to such an estimation one can use different methods, which will yield different results. The simplest way is to use the software log of the OCR engine, a file in which the software documents (amongst other things) whether a letter or a word has been recognised correctly according to the software's algorithm. While this can be used with other (often special)

²⁵ See <http://djvu.org/>

²⁶ See especially the chapter "3.6.3 Re-consider checking OCR output accuracy", below.

software and thus allow for the verification of a complete set of OCRed material, it is also of rather limited use. The reason for this is that the OCR software will give an estimation of how certain the recognition is according to that software's algorithm. This algorithm cannot realise any mistakes made because they are beyond the software's scope. For example: Many old font sets have an (alternative) 's', which looks very similar to an 'f' of that same font set. If the software has not (properly) been trained to recognise the difference it will produce an 'f' for every such 's'. The software log will give high confidence rates for each wrongly recognised letter and even the most advanced log analysis will not be able to realise the mistake.

The second method for estimating the correctness of OCR output is the human eye spot test. Human eye spot tests are done by comparing the corresponding digital images and fulltext of a random sample. This is much more time consuming than log analysis, but when carried out correctly it gives an accurate measurement of the correctness of the recognised text. Of course, this is only true for the tested sample, the result for that sample is then interpolated to get an estimation of the correctness for the whole set of OCRed text. Depending on the sample, the result of the spot test can be very close to or very far from the overall average of the whole set.

3.6.2 Letter count vs. word count

After deciding on the method for estimation, one has to decide what to count. One can compare either the ratio of incorrect to correct letters or the ratio of incorrect to correct words. The respective results may again be very different from each other.

In either method, it is important to agree on what counts as an error. One could for example, count every character (including blank spaces) that has been changed, added or left out.

For example: The word 'Lemberg' has been recognised as 'lern Berg'. In letter count, this would be counted as five mistakes: 1: 'l' for 'L', 2: 'r and 'n' for 'm', 3: one letter added, 4: blank space added, 5: 'B' for 'b'. Notice that the placement of 'r' and 'n' for 'm' counts as two mistakes!

In word count the same example would count as two mistakes. One, because the word has been wrongly recognised and two, because the software produced two words instead of one.

Currently, the letter count method is mostly used, because it produces the same difference in the average for each detected error. That is each detected error is counted as one error, regardless of its importance within the text. The problem with letter count is that it is impossible to make statements about searchability or readability from it.

The word count average, on the other hand, only changes if a new error also appears in a new word. That is to say, when two letters in a single word are recognised wrongly, the whole word still counts as a single error. If an error is counted, though, it usually changes the average much more drastically than it would in letter count, because there are fewer words in a text than there are letters.

While word count will give a much better idea of the searchability or readability of a text, it does not take into account importance of an error in the text. Thus an incorrectly recognised short and comparatively unimportant word like "to" will change the average as much as one

in a longer word like “specification” or a medium sized word like “budget”. Thus, the predictions about searchability or readability of a text made from word count are not very accurate either.

Only a very intricate method that would weigh the importance of each error in a given text could help here. There are now projects working on this problem, but there is as yet no software that does this and employing people to do it would not be practical.

3.6.3 Re-consider checking OCR output accuracy

Because of the problems with all methods described above and because the simple estimation of the percentage of errors in a text does not change the quality of current OCR software, libraries planning large scale digitisation projects should consider refraining from checking the quality of their OCR results on a regular basis. Even in smaller projects, where checking OCR results is more feasible, the amount of work put into this task should carefully be considered.

This said, at least at the beginning of a project the OCR output should be checked to a certain extent to make sure that the software has been trained for the right fonts, the proper types of documents and the correct (set of) languages.

Also, to get a simple overview of the consistency of the OCR output and to find typical problems, it may be a good idea to put the software's estimated correctness values into the OCR output file (ALTO for example already provides for this) or to keep it separately. A relatively simple script can then be used to monitor these values and to find obvious discrepancies. These can then be followed up to see where the problem is and what, if anything, can be done about it.

3.7 Post-processing

Post-processing for OCR'd historical text is still a fairly new topic. While it is being worked on, there are currently few ready-made, satisfying tools for this. OCR post-processing is at this point not proposed as a viable part of planning an OCR workflow, but rather a pointer to a field where planning for the future may be important.²⁷

3.7.1 Lexical correction

One field of post-processing that is already in use, but may become more interesting in the future, is lexical correction.

"ABBYY FineReader 9.0 supports 184 recognition languages, including 38 languages with dictionary support. For languages with dictionary support you may use the FineReader spell-checking system."²⁸

²⁷ The various methods of post-processing explained here are currently research topics in the IMPACT project (<http://www.impact-project.eu/>). IMPACT (IMproving Access to Text) is funded under the Seventh Framework Programme of the European Commission (FP7) (<http://cordis.europa.eu/fp7/ict>)

²⁸ [01]

As has been shown in the "Survey of Availability of Digitised Images for OCR"²⁹, ABBYY's FineReader is in one version or another, sometimes as part of a different programme or a group of programmes, the OCR software being used by all content providing TELplus partners. For this reason and seen without reference, the above statement may lead to the belief that the situation for lexical correction of OCR results may already be a fairly good one.

Looking closer at the list of the languages with dictionary support, one starts to get an idea of where the main problems lie. Even though those dictionaries support modern languages, there are already three varieties for Armenia: Eastern Armenian, Western Armenian, and Grabar. The explanation given is the following: "Grabar – is now used exclusively as the language of the clergy. The modern literary language has two main varieties – Eastern (Yerevan), spoken in Armenia and Western, spoken in Near East and Western Europe."³⁰ For German there are two varieties: new and old spelling. Out of the 38, five languages have more than one variety.

While today, we can expect standardisation for the majority of modern languages, historical texts will show large amounts regional and temporal differences in the spelling of common words as well as of place or personal names.

One of the best-known examples for this is the spelling of William Shakespeare's family name. David Kathman says that "Elizabethan spelling was very erratic by twentieth-century standards, though it was not (as is sometimes stated) totally without rules. Even the simplest proper names were spelled a variety of ways [...]".³¹ He finds no less than 26 different spellings of the name.

Because of the amount of varieties in spelling and because OCRing of historical text is still a fairly new field, there is as yet nearly no dictionary support for historical texts. This means that current projects must either rely on dictionaries that are not really fit for their purpose or produce dictionaries with which the software can be trained beforehand. This last is obviously only a choice if the software used allows for such training.

There are two approaches to supply dictionary support: The first is, of course, the creation of a lexicon of words and variations for a given period and area. The second is a list of so-called Named Entities (NE). An NE-list is a list of names of known organisations, persons, and geographical places and includes possible variations of these (e.g. all variations of Shakespeare's name).

Of course, neither kind is really possible for any kind of OCR project, as both can only be created with a very good knowledge of historical languages and therefore the help of linguists. Only a very limited NE list, which only names those persons and places that are certain to appear in the OCRred set, may be a feasible option, but even for the limited NE list cost and benefit should be analysed.

²⁹ [13]

³⁰ [01]

³¹ [10]

It can be hoped, though, that in the near future tools will be developed that make the integration of exciting lexica and NE lists. It may therefore be profitable to plan for projects to develop them, independent of current or planned OCR projects. This is especially true for languages with relatively small speaker groups, as these will most probably not become the direct focus of such tool developments.

3.7.2 Collaborative correction

A second field of OCR post-processing, which is also already being explored, is that of collaborative correction. The idea here is to allow the library users direct access to the fulltext and provide them with tools to correct OCR errors. Here as with the dictionaries, there is not yet a big market for tools or applications, but development of such applications is in progress.

One point that seems important in any collaborative correction is that they will take time and effort to set up, but that they will enable the library to put less effort in preparing their material for OCR. A dedicated community will have to be built up around these projects. They will demand, as all such communities, a certain amount of maintenance and ministration, but they will most likely add to the library's credibility and connect it more closely to its user base in the long run.

The following are two examples of national libraries have tried to involve their users in the process of collaborative text correction.

3.7.2.1 Distributed Proofreaders

The National Library of Portugal had a project harnessing the possibilities of the Distributed Proofreaders (DP).³² DP is a sub-project of Project Gutenberg,³³ which provides a web-based method for the collaborative conversion of Public Domain books into e-books.

The BNP started this project by translating the DP web site and documentation into Portuguese and announcing this on their Digital Library web site. A project coordinator was selected and a workflow was put in place.

Usually the coordinator selects a book and contacts the BNP in order to get the images to work on. Not every book may be "proofreadable". There are several points that need to be taken into account:

The book needs to be in Public Domain in Portugal and in the USA. Under Portuguese law, the book has to be from an author who has died at least 70 years ago, and the edition used may not have been published after the author's death. For the USA, the only thing that matters is the date of publication. Foreign books have to have been published before 1922.

The images need to have at least 300 dpi to obtain good OCR text. Otherwise, it is preferable simply to have the book keyed.

³² [05]

³³ [06]

All finished Portuguese books are to be found at the Project Gutenberg Webpage. Those, whose source images came from the BNP, have it announced at the beginning of the e-book. There is usually an HTML and an ISO-8859³⁴ version for each book. DP does not keep the original images, nor does the Gutenberg Project give access to them.

The main advantage of the DP concept is that books are proofed at least 5 times (by five different people). The Portuguese books then have a final revision by the project coordinator to assure consistency. This means that the library that uses DP has a lot of control over the results and will be sure to get high quality products.

The problem with this approach is that there is no way to know how long a single book will take to go through the whole process. The library will have to put some effort into creating a dedicated community around the project. This will, of course be a good deal easier for languages with large speaker bases, but should be possible for any language.

Also, the choice of material that may be corrected through the DP process is somewhat limited. This is due to the fact that DP only corrects books, but also, because those books will have to be of general interest to attract proofreaders.

It has to be kept in mind that the library has to be willing to have an e-book version of one of their book sitting in the "shelves" of the Gutenberg Project, which will mean to direct a certain amount of traffic away from the library's website.

There is a certain amount of effort involved in coordination and in choosing the material for the project. The library now has e-books that it can present its readers with. However, to provide this text in such a way that the images are connected to it, either via a METS/ALTO setup or in PDFs with text under image or in DjVU files, would take a whole lot of additional effort.

The National Library of Portugal has stopped the project because of some restructuring processes. The former library coordinator is now leading the project independently from the library.³⁵

3.7.2.2 Provision of a correction interface on the library website.

An alternative to passing collaborative OCR correction to an external organisation is the provision of an interface on the libraries own website. There are already a number of projects that work on such an interface. One current example is the National Library of Australia's (NLA) Australian Newspaper website, which is currently in a beta status.³⁶

The NLA's own argumentation for the interface they provide closely mirrors those of similar projects:

³⁴ A set of ISO/IEC standards for 8-bit character encodings, which includes the well known Latin-1, Latin-2 etc. sub- standards (ISO-8859-1, ISO-8859-2 etc.).

³⁵ The new project site can be found here [16]

³⁶ [02]

"The quality of the OCR text may be poor, but the actual page image readable to the human eye. Users may wish to correct the OCR data. If a word is wrong in the OCR it will not be retrieved in the search results but if it is corrected it will. Making corrections to people's names in particular would enhance accuracy and usefulness of name searching for genealogists. Users may wish to monitor and let other people view how many corrections they make and become 'OCR Correction Kings'. OCR corrections could be moderated in a similar way to Wikipedia or the Library could trust the community and let OCR corrections be a freely flowing process, monitored by the community itself."³⁷

As the DP example shows,³⁸ there are quite a few users that are indeed willing to put some effort into correcting OCR'd text. Discussions among TELplus partners have shown some tendency to having a moderated process rather than one in which the community members are left to their own device. This was especially true for those participants who feared vandalism³⁹. It will be interesting to follow the example of the NLA and see what their experience will be.

There are, of course, also limits to what may be achieved with this model of community involvement, though there are no studies of such projects yet. The NLA provides the interface only for the correction of OCR'd newspapers. This is a good step as many libraries currently OCR newspapers, but it will be important to see whether a community that has been build around the correction of book or newspaper texts can be inspired to also provide their combined expertise to other material which may be less interesting to larger groups of people.

The NLA provides a good variation⁴⁰ on the warnings described in the chapter "Choice of material for OCR" above:

The text in the left panel has been electronically translated by a computer. Computers are not as good at reading as humans, and often make mistakes.

You can help correct mistakes in articles by moving your cursor over a line and clicking "Help fix this text".

By helping to fix this text, you are making it easier to search and a better resource for everyone!

While the interface is very intuitive and allows for easy access to the text and for the correction of it on the fly, there is as yet no great advertisement for this part of the project.

³⁷ [08] page 1.

³⁸ On 5 August 2008 641 users were registered to have been active in the past twenty-four hours; 1.362 in the past 7 days; and 2.800 in the past 30 days.

³⁹ Any addition, removal, or change of content made in the attempt to change the original information or just as a way of saying "I was here".

⁴⁰ This is actually a pop-up window that comes up when the reader clicks the "Why may this text have mistakes?" or the "Help fix this text!" link on an individual newspaper display page. It may be found here: <http://ndpbeta.nla.gov.au/static/ndp/oxideDesign/ocrHelp.html>

There is no direct link from the homepage and no community page. These would be a good idea, but the project is currently only in beta and thus such things may follow as it progresses.

3.8 Cooperation

As has been seen above, the main point in all OCR projects is always holding the balance between the effort put into choosing and preparing the input for OCR on the one hand, and the effectiveness in quality and quantity of the output on the other. A suggestion that is usually made in this context is that such projects should look for partners who either share the load of the work or finance it. This means either finding other libraries that have similar interests or finding partners for so called "private-public partnerships" (e.g. those between libraries and Google).

3.8.1 Inter-library cooperation

There are, of course, quite a number of fields in which libraries and more specifically national libraries of different countries can cooperate. This section is only concerned with suggestions for cooperation in digitisation and OCRing.

The first suggestion is aimed specifically at smaller libraries or libraries with smaller digitisation and OCR programmes. These libraries should consider planning their programmes in cooperation with libraries in similar situations, setting up joint tenders for OCR. This would lead to larger budgets and better leverage in the negotiations with potential contractors. Also, some potential contractors are only interested in taking projects when they have a certain size.

Another possibility would be, if one of the potential partners already does their work in-house, has free capacities or the potential to broaden their capacities. They could act as a contractor to the others. This is in fact what the National and University Library of Iceland does for the national libraries of the Faro Islands and Greenland.

This example also shows one of the challenges of such cooperation. All three languages involved are rather small languages, so support of these languages in OCR software can be expected to be limited.

So while such a cooperation possibly makes the digitisation of text source of the smaller languages feasible at all, it also puts additional challenges on the largest partner in the partnership. Such scenarios have to be considered before beginning such a joint project.

Another field of inter-library cooperation opens up for national libraries of countries with common histories. Typical examples would be Middle-European countries (e.g. Austria) and their former dependents, the Scandinavian countries, the countries of former Yugoslavia or other such combinations. Wherever there was movement of borders or groups lived on both sides of a border, such cooperation suggests itself.

A first step in this direction was made, when the National and University Library of Slovenia (NUK) decided to digitise the "Laibacher Zeitung". Laibach was the official name of

Ljubljana⁴¹ when Slovenia was part of Austria. Thus it was logical to assume that the Austrian National Library also had an interest in the digitisation of the newspaper. During the ensuing negotiations it was discovered that both sides had almost the same stock. NUK decided to digitise only its own stock rather than coordinating the delivery of a few issues that would not add much value to the outcome of the project. However, both libraries are interested in future projects, though none is currently planned.

3.8.2 Private-public partnerships

One possibility of financing larger digitisation and OCR projects is that of so-called private-public partnerships. In these partnerships, a commercial entity (e.g. Google) finances the digitisation of collections of a non-commercial entity (in this case a national or other library).

It is obvious that the commercial entity will only enter into such a partnership if sufficient, direct or indirect (e.g. via advertisement), profit can be expected as the outcome of it. For this reason, the entity will aim at restricting access to the digitised material in a way that makes it the main, if not only access point to this material. The non-commercial entity, on the other hand, needs to keep its mission (which is usually open access to the material it owns) in mind.

There have been many discussions focussing mainly on the position of the non-commercial entity. Therefore, and because this is not the place to repeat the arguments of others at length, only those points that are most important to keep in mind will be collected here.⁴²

Kaufman and Ubois⁴³ have made the point that both sides should be aware of the needs of the other side. They urge libraries that consider such partnerships to define "walk away points" before they enter into negotiations and be prepared to negotiate extensively. To this end, libraries should be aware that bring more into such partnerships than just books. They bring their experience in collecting, preserving and giving access to these books, thus bringing more than just the material basis to the partnership.

Kaufman and Ubois give a list of requirements that libraries should insist on to maximise their gain from a private-public partnership⁴⁴:

- Limited confidentiality – Libraries should be sensitive to private partner needs to protect business and technology secrets, but insist on their own right to discuss aspects pertaining to their broader community. These deals involve some of the most complex decisions libraries will face, they can be improved through consultations with others, and libraries do not want to be in the position of having to refuse advice to peers who seek their guidance.

⁴¹ Slovenia's capital.

⁴² See <http://www.oclc.org/programs/ourwork/collectivecoll/harmonization/massdigresourcelist.htm> for a selection of materials about mass digitization, public/private partnerships and the agreements that govern them, compiled for the OCLC's "Harmonizing Digitization program".

⁴³ See [11] This document gives a very good overview of the subject, because it gives a good idea of what to expect from the private partner. I also collects the experiences of many from participants of the early large-scale digitisation programmes.

⁴⁴ [11] page 2

- More complete deliverables – Librarians must have input into the specifications of quality and formats and be clear about exactly what they will receive. They must ensure that they will own those deliverables.
- More open access – Librarians should preserve their right to provide unrestricted access to users. In particular, they should avoid contract terms that make it difficult or impossible to offer scholars the kinds of functionality, including automated or bulk access to collections, that can support innovative research and will allow the development of new applications.
- Less restricted distribution – Librarians should preserve the right to combine parts or all of their digitized content with collections at other institutions or nonprofit organizations.
- Responsible treatment of usage data – Librarians should ensure that users' privacy is protected, even while drawing on usage data to enhance services to users.

If the above terms cannot be secured, then the consequences of compromises should be fully understood. This point, however, should be a requirement:

- Limited duration and survivability – Restrictions on ownership, access, and distribution should not survive termination of the agreement.

Kaufman and Ubois make this distinction between that last point and the other. While this is probably meant as a kind of "limited damage clause", libraries should not give up their rights to act do what they are meant to do, i.e. give access to and preserve their holdings. This should include digital representations.

A second point a library should consider before entering into such a partnership is whether it has the capacities to actually provide access to the digitised material. In his blog entry from August 25, 2008, Roy Tennant claims that "The University of Michigan Library led the way in putting the books online that Google was digitizing from their collection, and to this day remains one of the few (the only?) to have made the effort."⁴⁵ This means that access to the rest of those digitised collections is only possible through the private partners and thus only with the limitations they may want to place on the access.

3.9 Summary

These good practice recommendations have been written to give an overview of the points that have to be considered when planning an OCR project. Many of these points, like the choice of material, will already affect the digitisation workflow, others, e.g. cooperation with other libraries, may only be considered after a library has had some experience with such projects.

⁴⁵ [22]

Because of the differences in planning different types of projects, these recommendations cannot act as a checklist, they will more likely be a list of points to be aware of and suggestions of how to deal with these points once they become important for a project.

On some points, the suggestions have had to remain vague, as little or no large-scale research has yet been published on these points. It is, however, to be hoped that research like that done in the IMPACT project⁴⁶ will change this in the future.

As many of the points raised and solutions given here are as much a matter of experience as of opinion, that these recommendations are taken as a starting point for further discussion rather than as the attempt to speak the last word on the subject.

⁴⁶ IMPACT (IMproving Access to Text / <http://www.impact-project.eu/>) is funded under the Seventh Framework Programme of the European Commission (FP7) (<http://cordis.europa.eu/fp7/ict>)

References

- [01] ABBYY: "[Technical Specifications Recognition Languages - ABBYY Fine Reader Professional Edition](#)". (last accessed 2008-08-18)
- [02] Australian National Library: "[Australian Newspapers beta](#)". (last accessed 2008-08-25)
- [03] Carney, Bill (William): "[Automating registration of digital preservation copies - the place of registries in the digitisation workflow](#)" Presentation given at the LIBER-EBLIDA Workshop on Digitisation at the Royal Library, Copenhagen, Denmark 24 October 2007. (last accessed 2008-06-03)
- [04] Carney, Bill (William): "[Automating registration of digital preservation copies: the place of registries in the digitisation workflow](#)" In: Liber Quarterly, [Volume 18 \(2008\), No. 1](#). (last accessed 2008-08-18)
- [05] Distributed Proofreaders: "[DP homepage](#)". (last accessed 2008-08-25)
- [06] Project Gutenberg: "[Project Gutenberg](#)". (last accessed 2008-08-18)
- [07] Holley, Rose: "[Optical Character Recognition \(OCR\) on Newspapers – An Overview](#)". 2007. (last accessed 2008-06-03)
- [08] Holley, Rose: "[User Interaction Ideas: What if there were no boundaries - how would users want to interact with newspaper data, other users, and add value to the data and the service?](#)". 2007. (last accessed 2008-06-03)
- [09] IMPACT project: "[IMPACT project homepage](#)". (last accessed 2008-08-18)
- [10] Kathman, David: "[The Spelling and Pronunciation of Shakespeare's Name](#)" (last accessed 2008-08-18)
- [11] Kaufman, Peter B. and Ubois, Jeff: "[Good Terms - Improving Commercial-Noncommercial Partnerships for Mass Digitization: A Report Prepared by Intelligent Television for RLG Programs, OCLC Programs and Research](#)". In: D-Lib Magazine, Nov./Dec. 2007 [Volume 13 Number 11/12](#). (last accessed 2008-08-25)
- [12] King, Ed: "[British Library Digitisation: access and copyright](#)". 2008. (last accessed 2008-09-30)
- [13] Korb, Joachim: "[Survey of Availability of Digitised Images for OCR](#)". 2008. (last accessed 2008-08-18)
- [14] Mark, Timothy: "[National and International Library Collaboration: Necessity, Advantages](#)" In: Liber Quarterly, [Volume 17 \(2007\), No. 3/4](#). (last accessed 2008-08-18)
- [15] National Library of New Zealand: "[Papers Past – Technology](#)". (last accessed 2008-08-18)
- [16] Página a Página: "[Página a Página homepage](#)". (last accessed 2008-08-18)
- [17] Riger, Oya Y.: "[Preservation in the Age of Large-Scale Digitization - A White Paper](#)". 2008. (last accessed 2008-06-03)

- [18] Schwartz, Werner: "[EROMM and the Registry of Digital Masters](#)" Presentation given at at the LIBER-EBLIDA Workshop on Digitisation at the Royal Library, Copenhagen, Denmark 24 October 2007. (last accessed 2008-06-03)
- [19] Schwartz, Werner: "[EROMM and the Registry of Digital Masters](#)" In: Liber Quarterly, [Volume 18 \(2008\), No. 1](#). (last accessed 2008-08-18)
- [20] Stockmann, Ralf: "[Central registry for digitized objects: Linking production and bibliographic control](#)" Presentation given at at the LIBER-EBLIDA Workshop on Digitisation at the Royal Library, Copenhagen, Denmark 24 October 2007. (last accessed 2008-06-03)
- [21] Tanner, Simon: "[Deciding whether Optical Character Recognition is feasible](#)". 2004. (last accessed 2008-07-11)
- [22] Tennent, Roy: "[Hathi Trust Fund](#)". In: [Tennant: Digital Libraries. Roy Tennant's news and views on digital libraries](#). (blog entry of 2008-08-25) (last accessed 2008-08-26)
- [23] Times online: "[What is the Times Archive?](#)" (last accessed 2008-06-03)
- [24] Verheusen, Astid: "[Mass digitisation?](#)" Presentation given at at the LIBER-EBLIDA Workshop on Digitisation at the Royal Library, Copenhagen, Denmark 25 October 2007. (last accessed 2008-06-03)
- [25] Verheusen, Astrid: "[Mass Digitisation by Libraries: Issues concerning Organisation, Quality and Efficiency](#)" In: Liber Quarterly, [Volume 18 \(2008\), No. 1](#). (last accessed 2008-08-18)