

ECP-2006-DILI-510003

TELplus

Survey of Availability of Digitised Images for OCR

Deliverable number	<i>D-1.1</i>
Dissemination level	<i>Public</i>
Delivery date	<i>24 January 2008</i>
Status	<i>Final 1.1</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. Introduction	5
1.1. <i>The aim and scope of this document</i>	<i>5</i>
1.2. <i>The structure of this document</i>	<i>5</i>
2. The material to be OCR-ed for WP 1.....	6
2.1. <i>Agreed amounts / promised amounts</i>	<i>6</i>
2.2. <i>Types of material to be OCR-ed for WP 1</i>	<i>8</i>
3. OCR: Software, accuracy, in-house OCR-ing / out-sourcing	10
3.1. <i>OCR production</i>	<i>10</i>
3.2. <i>OCR accuracy</i>	<i>11</i>
3.3. <i>In-house / out-sourcing</i>	<i>13</i>
3.4. <i>Learning, standardising, and benchmarking</i>	<i>14</i>
4. OAI compliance	14
5. Summary / outlook	16

Executive Summary

This is the first deliverable of TELplus workpackage 1, “Making Searchable Digitised Images via OCR”. It presents the results of Task 1.1 the “Survey of availability of digitised images for OCR”. Task 1.1 is one of a pair of surveys designed to gather information needed as a starting point for all other Tasks in WP 1.

TELplus WP 1’s goal is to use digitised images of historical text for the creation of full texts via OCR. The full text produced will be used for detailed searching in existing digital libraries; thus, a considerable added value will enrich the existing data. In this way, the partners will provide better access to more than 20 million pages of digitised documents, mostly historical periodicals. All the OCR-processed collections will be searchable via The European Library, preferably through OAI-PMH protocol. Best OCR practices will be mapped as well as structured electronic delivery of full texts.

The present survey asked partners to describe the material they offer as their share of TELplus WP 1, their experience of OCR, and whether their respective collections are currently accessible for The European Library via the OAI-PMH (OAI protocol for metadata harvesting; <http://www.openarchives.org/OAI/openarchivesprotocol.html>).

This last point is especially important in planning Task 1.5 in which the new material will be made accessible through The European Library. It was hoped that the OAI-PMH could play an important role as a means of providing information about the collections and the new full-text, to The European Library. This information would have been used by The European Library to harvest and then index the full text, but the survey showed that the majority of collections are currently not accessible via the OAI protocol. This will have to be taken into account and remedied by WP 2.

The outcome of the answers to the other questions of the survey was more encouraging. The partners have all confirmed the number of pages they originally agreed to OCR. Some are even confident to be able to provide more than their share. The material provided by them is comprehensive both in timescale and in geographical spread and language variety. We will have full-text of material dating from between 1500 and 2002, from sources geographically as far apart as the Mediterranean, Greenland, or even South America, and printed in no less than 30 languages. Thus we can now be certain that TELplus will succeed in helping The European Library in getting the critical mass of digital material it needs to fulfil its mission.

The third group of questions was aimed at providing a picture of how experienced in OCR matters TELplus WP 1 partners are. The outcome of this was much as was expected. WP 1 brings together very experienced partners like the French National Library and partners that have just started expanding their work into the field of OCR-ing digitised material. One of the important points of WP 1 was bringing those two groups together to provide a platform for exchange of information and development of new expertise.

Thus Task 1.1 has managed to provide a basis for all other Tasks in WP 1, as well as important information for WP-2 and for The European Library. The next step will be a workshop in Paris in January 2008, which, together with the results of this survey, will form the groundwork for the second survey in Task 1.2, “Survey of Existing OCR Approaches”,

and for the partners OCR implementation plans that are the main object of Task 1.3, “Identification of Concrete Materials for OCR; OCR specification, implementation plans and tenders”. This, with this document, lays the groundwork for the whole OCR workpackage of TELplus and will be part of the foundation of its success.

1. Introduction

Many libraries have digitised content that is not searchable in detail, as no descriptive metadata are used on lower levels of the document structure and 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 text for the creation of full texts via OCR. The full text produced will be used for detailed searching in existing digital libraries; thus, a considerable added value will enrich the existing data.

This full-text material will be delivered to The European Library by 14 different National Libraries, each with their own background and experience. To get an overview of the material and situations of these libraries two surveys have been planned for WP 1. This first survey, the results of which this document describes, gives an overview of the different approaches to OCR the partner libraries have taken and a first insight into the challenges which arise when dealing with historical and diverse material with different fonts, quality of original material, and even languages the material was printed in. It will also give a first idea of what the staff of The European Library will have to deal with when they integrate the TELplus results into The European Library’s services.

1.1. The aim and scope of this document

This document’s aim is to give an overview of the full-text material that is to be expected as the outcome of TELplus WP 1. It is meant to deliver the basis for Task 1.3, in which the individual implementation plans for WP 1 partner libraries will be drawn up. It will also give a first impression of the common challenges WP 1 partners are going to face and it will provide a starting point for the planning of Task 1.5 which will integrate the new material into The European Library.

While many important providers of digitised and OCR-ed material in Europe are also partners in TELplus WP 1, this document is not designed to provide an overview of the situation of OCR in Europe. However, given the differences in experience and organisational structure of the TELplus partner organisations, many of the points raised in this document may well provide a good basis of discussion even outside of TELplus.

1.2. The structure of this document

First an overview of the material to be OCR-ed for TELplus will be given. The amount each library is going to deliver will be established and compared against the amount agreed upon in the Description of Work. The material will be described by type, location, timescale, and language.

Secondly the differing procedures of OCR-ing at the respective libraries will be described. Special focus will be put on the software being used and the accuracy of the OCR results. A differentiation between in-house and out-sourced OCR production will be made.

This document will also provide information about current OAI-PMH (OAI protocol for metadata harvesting; <http://www.openarchives.org/OAI/openarchivesprotocol.html>) accessibility of those collections that will be OCR-ed for WP 1.

Finally, an outlook over the next steps in WP 1 will be given.

2. The material to be OCR-ed for WP 1

Most important in WP 1's work for The European Library is, of course, the full-text material that will be produced. Besides the final amount of OCR-ed pages, type and context of the original material are crucial for the future relevance of the produced full-text. This last point will be determined by a good spread of geographical origin and language and timescale. Accordingly this chapter will give an overview of numbers and background.

2.1. Agreed amounts / promised amounts

The table below shows how the originally agreed number of 22.540.000 OCRed pages is allocated to the content providing partners. All content providers have confirmed their dedication to produce at least that amount (either at the TELplus kick-off meeting in Tallinn or, in the case of those excused for that meeting, in later communication). The originally agreed numbers are found in the "agreed number of pages" column, while the numbers partners now predict they will deliver are to be found, sorted by type of material, in the "predicted number of pages" column.

In their answers to the survey, the partners were asked to indicate the types and respective amounts of material they were planning to OCR for the TELplus project. As can be seen in the table, a number of partners now believe they will be able to produce even more than originally agreed. In other cases the given figures do not add up to the agreed amount or figures are not given by document type. The National Library of Norway does not give us figures at all.

These discrepancies can be explained by the way the OCR results are being produced at the respective libraries. In the case of the National Library of Norway, for example, digitisation and OCR are done in sequence so that, where possible, all material that is digitised is then also OCR-ed. In other cases a workflow for OCR has already been established. Here also the decision of what to OCR is being decided on in the course of the workflow. In both scenarios the partners know from experience how much full-text they usually produce in a given time without necessarily being able to predict the material that will be OCR-ed in half a year or a year from now.

Some of the partners have no or little experience with OCR and are expecting to significantly improve their knowledge by participating in the project. As a result, at this stage the plans of some libraries are somewhat tentative. They know which collections they are likely to use and how many pages they want to produce over the course of the program, but may be less certain

of results with specific types of material. For example this is the case with the National Library of Estonia. They are certain they want to OCR newspapers and journals, but will decide on specific amounts according to results.

If, as is to be expected, no great changes are made, more than half of the pages OCR-ed for TELplus will be from newspapers. Currently book pages make up about 10% of all pages to be OCR-ed, but this number is very indefinite, because of the uncertainties described above. All other amounts are even more difficult to predict, but the third choice of type of material for OCR will probably be journals and periodicals.

Table 1: Promised and predicted amounts of TELplus OCR-ed material

National Library	Agreed number of pages	Material	Predicted number of pages
Austria	500.000	newspapers	150.000
		governmental material	411.000
Czech Republic	3.400.000	books	2.400.000
		newspapers	1.000.000
Estonia	200.000	newspapers	at least 160.000
		journals	at least 10.000
France	7.000.000	books	7.000.000
		periodicals	
Hungary	200.000	periodicals	4.703
		newspapers	45.014
		journals	111.362
		books	5.000
		monographs	33.921
		pamphlets	1.600
Iceland	2.800.000	journals	260.000
		newspapers	2.190.000
Latvia	100.000	books	75.000
		newspapers	25.000
Lithuania	100.000	newspapers	100.000
Norway	1.600.000	books	numbers depend on the results of current digitisation efforts
		journals	
Poland	420.000	books	350.000
		newspapers	70.000
Slovak Republic	200.000	newspapers	> 200.000 total

		journals	
Slovenia	320.000	newspaper	70.000
		journals	200.000
		books	50.000
Spain	5.500.000	newspapers	5.000.000
		books	> 100.000
Sweden	200.000	newspapers	320.000
		journals	
		books	
		printed ephemera	
Total	22.540.000		

2.2. Types of material to be OCR-ed for WP 1

So what type of material will actually become accessible through the work done for TELplus? As has been pointed out above, the majority will be from newspapers. Most National Libraries, as is to be expected, use material that is originally from their own country and in their own language. Also not surprising is the fact that some countries use material from countries they were once part of or linked to. Most notably, Hungary has material from 6 different countries and in 4 different languages (including English, from modern text books, and Latin). Similarly, the National Library of Poland is OCR-ing texts not only in Polish, but also from up to 6 “minority languages”. By including newspapers and journals from the Farø Islands and Greenland, the National Library of Iceland extends the area where the material comes from almost across the Atlantic. Besides its own, the National Library of Iceland thus also provides us with digitised and OCR-ed texts in two very rare languages.

Just as the origin of the TELplus material is spread over a vast area, geographically, so it is spread over several centuries in its timescale. Slovenia will contribute the oldest book, which will be from 1500, while the youngest Icelandic newspaper will only be 6 years old. The majority of the material will be from the time between the early 19th to the mid 20th century. Access to later material is often encumbered by copyright issues. Often this kind of material is only digitised when the library can be certain of its legal status.

Thus we can be sure that the material The European Library gains through TELplus will be relevant and interesting to a large and divers audience.

Table 2: Geographic background, timescale, and languages of the TELplus OCR material

National Library	Material	Timescale	Geographic coverage	Languages
------------------	----------	-----------	---------------------	-----------

Austria	newspapers	1900 – 1925	Austria	German
	governmental material	1862 – 1918		
Czech Republic	books	1800 – 1930	Czech Republic	Czech, German
	newspapers	1800 – 1989		
Estonia	newspapers	1821 – 1918	Estonia	Estonian
	journals	1890 – 1940	Estonia	Estonian
France	books	1650 – 1930	France	French (some others)
	periodicals			
Hungary	periodicals	1721 – 1796	Hungary, Slovakia, Romania, Austria	Latin, Hungarian
	newspapers	1838 – 1918	Hungary	Hungarian
	journals	1867 – 1944	Hungary, Serbia	Hungarian
	books	1590 – 1726	Hungary, Poland, Slovakia, Austria, Germany	Latin, Hungarian, German
	monographs	1862 – 1992	Hungary	Latin, English, German
	pamphlets	1956	Hungary	Hungarian
Iceland	journals	1773 – 1920	Iceland	Icelandic
	newspapers	1913 – 2002		
	journals	1890 – 2001	Farø Islands	Farøese
	newspapers	1861 – 1999	Greenland	Greenlandic
	journals			
Latvia	books	1900 – 1945	Latvia, Estonia, Germany	German, Latvian
	newspapers	1900 – 1952	Latvia	Latvian, German
Lithuania	newspapers	1904 – 1940	Lithuania	Lithuanian
Norway	books	mostly books by authors dead for more than 70 years	Norway	Norwegian (others)
	journals			
Poland	books	before 1939	Poland	Polish
	newspapers	1918 – 1939	Poland	Polish (German, Czech, Ukrainian, Byelorussian, Yiddish)
Slovak Republic	newspapers	before 1918	Slovak Republic (others?)	Slovak, Hungarian, German
	journals			
Slovenia	newspaper	1888 – 1944	Slovenia	Slovenian
	journals	1843 – 1902		
	books	1500 – 1945		
Spain	newspapers	17 th – 19 th century	Spain and Spanish speaking Countries in South America	Spanish
	books			
Sweden	newspapers	still to be decided	Sweden	Swedish
	journals			

	books			
	printed ephemera			

3. OCR: Software, accuracy, in-house OCR-ing / out-sourcing

Within TELplus WP 1 we have a number of important points in connection with the way the OCR work will be done. About half of the contributing partners will produce their contribution via a sub-contracted service provider. So sub-contracting and especially the connected tendering procedures are special focus points in our preparation.

Secondly, OCR-ing, especially that of historical material, is still a very new subject. Problems like deterioration of original material and special fonts are obstacles that need to be overcome to achieve good quality full-text of digitised materials. As opposed to modern office documents, which are usually printed in standard “Antiqua” type fonts (like the well known “Times New Roman”, “Courier”, and “Arial” fonts), the historical documents WP 1 is concerned with are often printed in old fonts that are difficult to read, even for most modern readers. Due to the fact that they were mastered on handwriting using a pen with a nib, these fonts (especially gothic, black letter or Fraktur fonts) are made up of dots and lines. Many of these are so similar that even the smallest smudges may lead to confusion. Similar problems occur with diacritics like, for example, the French accents (e.g. é and è).

Given the quality of the original material that was scanned – brittle, dirty, well used paper – these problems are very common. Often whole parts of the text are unreadable or even missing. Here, where the human reader may successfully guess the missing passages, today’s software is still lacking capabilities that match human guesswork. Especially for languages with few native speakers, there are no or insufficient dictionaries to aid this work. Even languages that have a large speaker base have been subject to change or have large numbers of dialects. Thus standard dictionaries are of little help. Currently next to no adequate historical dictionary exist.

Also, as the subject of OCR-ing historical material is so new, methods of establishing the quality of OCR are not yet standardised. This chapter is designed to give an overview of how WP 1 partners have tried to overcome OCR challenges.

3.1. OCR production

The range of OCR software fitted to the needs of libraries is rather limited. The fact that all partners with any experience in OCR-ing use or have used some version of Abbyy’s FineReader (<http://finereader.abbyy.com/>) is an indicator of this. Besides versions 5.0 to 8.0, versions XIX (for Black Letter or Gothic (Fraktur) fonts) and the Recognition Server have been used.

Those libraries or subcontracted service providers that do not use Abbyy’s products directly have used some sort of derivative. The National Library of Norway, for example is 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 was based on a combination of OCR products – FineReader among them.

Accordingly the results of full-text production are comparable and depend more on the choice of original material than on the quality of the software. The only exception to this is the Martynas Mažvydas National Library of Lithuania. Here the results of OCR-ing are revised by hand, giving almost perfect results. This practice is not feasible for most libraries, though.

Table 3: OCR software in use by WP 1 partners

National Library	OCR - software
Austria	docWorks for tests only
Czech Republic	FineReader 5.0 , 6.0, 8.0
Estonia	FR 8.0, XIX, Abbyy Recognition Server
France	Service provider's own software + FR 7.0, OmniPage (http://www.nuance.com/omnipage/), ReadIris (http://www.irislink.com/c2-532-189/OCR-Software---Product-list.aspx), EasyReader (discontinued)
Hungary	FineReader 6.0-8.0 (9.0)
Iceland	FineReader 7.0
Latvia	no OCR-ing done yet
Lithuania	FineReader 8.0 + manual revision
Norway	docWorks
Poland	FineReader 8.0
Slovak Republic	FineReader 0.7-0.8
Slovenia	FineReader 8.0 pro / FineReader development versions
Spain	FineReader 8.0
Sweden	FineReader 7.0

3.2. OCR accuracy

Given the similarity of the material and the software to be used, one could suppose that the quality of most partners' OCR results should be in a similar range. The table below shows a very different picture.

The main reason for this is the way OCR errors are established. Almost half of the partners that already have some experience in OCR-ing simply use the software log, a file in which the software documents whether a letter has been read correctly according to the software's algorithm. Most of the others do human eye spot testing, in which a random sample is checked by someone comparing digital image to full-text. Both methods have their problems and, as the experience of the National Library of Poland shows, the results of both methods differ drastically. While the software log analysis has 90% of the original text recognised

correctly, human eye spot test established only a 60-70% accuracy (for the purposes of TELplusWP 1, “accurate” means that 95% of the text was recognised correctly!).

Only the National Library of Lithuania and the National Library of Poland estimate their results as 99% accurate, but this is after OCR results were revised manually. As has been pointed out above, this method is simply not feasible for most libraries.

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. All WP 1 partners have decided to count letters. The National Library of Iceland additionally counts words, but seems to get a slightly lower average that way, which is interesting as most literature on this subject suggests the opposite. The explanation given here usually is that one word may be made up of more than one misread letter while in letter count every misread letter would also count as a mistake.

The advocates of word count over letter count maintain that it is the words that are important to the full-text, while the proponents in favour of letter count hold that it gives a more appropriate indication of the quality of the software.

Table 4: Accuracy of OCR results and methods of establishing it

National Library	Material	Accuracy \geq 95%	Method establishing accuracy
Austria	newspapers	not established	
	governmental material		
Czech Republic	books	estimated 15% over all	Letter count, human eye spot test
	newspapers		
Estonia	newspapers	estimated 10% over all	Letter count, software log
	journals		
France	books	6% for most material	Letter count, human eye spot test
	periodicals	100% for newspapers	
Hungary	periodicals	not established	Letter count, software log
	newspapers	50%	
	journals	70%	
	books	not established	
	monographs	70%	

	pamphlets	not established	
Iceland	journals	estimated 80% over all	Letter and word counts, human eye spot test
	newspapers		
Latvia	books	not established	
	newspapers		
Lithuania	newspapers	estimated 99% for corrected text	Manual revision
Norway	books	not established	
	journals		
Poland	books	90% software; 60-70% human eye spot test; estimated 99% for corrected text	Letter count, software log and human eye spot test; manual revision
	newspapers		
Slovak Republic	newspapers	50% over all	Letter count, Software log
	journals		
Slovenia	newspaper	95% over all	Letter count, human eye spot test (spell checking planned)
	journals		
	books		
Spain	newspapers	not established	
	books		
Sweden	newspapers	estimated 85% over all; for Antiqua 90%	Letter count, human eye spot test
	journals		
	books		
	printed ephemera		

3.3. In-house / out-sourcing

OCR-ing, as digitisation, is a demanding practice that needs dedicated machines and personal. Some libraries do not have the means to OCR in-house and so have give that work to a sub-contractor.

Of the 14 partners in WP 1 seven - namely the National Libraries of Austria, The Czech Republic, France, Hungary, Latvia, Lithuania, and Spain - have a budget to out-source their OCR contribution.

The other seven content providing partners do the OCR-ing in-house or do the sub-contracting on their own account. Some of those partners already have experience in out-sourcing their OCR work – namely the National Libraries of Norway and Slovenia.

3.4. Learning, standardising, and benchmarking

One of the important effects of TELplus WP 1 will be the exchange of information on the topic of OCR. Especially in connection with the workshop hosted by the French National Library in January 2008 in Paris, this exchange will help develop a set of best practices in OCR.

Up to now each Library that started an OCR project had to, more or less, re-invent the wheel. They had to come up with their own specific solution. WP 1 partners, and through them hopefully others, will have a chance to learn from the experience of others.

As has been noted above, some of the partners in WP 1 have little or no experience in OCR-ing. With the National Libraries of the Czech Republic, France, Iceland, Norway, and Spain, we have partners in WP 1 that have already OCR-ed up to several million pages. Others like the National and University Library of Slovenia have OCR-ed a large variety of materials. Thus a large spectrum of experience is brought together in this workpackage and all participants will have the chance to learn. Obviously, there will be more learning done by some and more dissemination of knowledge by others, but it is to be expected that even the partners in France and Spain, who have the largest amounts of already OCR-ed pages, will be able to gain from this project. The two surveys in this WP will give us the information we need to maximise this gain.

4. OAI compliance

One of the most important questions of WP 1 is how the newly produced full-text will be accessible through The European Library. One idea is to provide information about the file types in which the material is presented and the location within the directory of the content management system of the library where it may be accessed for indexing via the OAI-PMH, which is the The European Library's preferred exchange protocol for metadata.

From the table below it is quite apparent that this is currently not a practicable solution, as the majority of collections concerned are not accessible via OAI-PMH. At the moment only metadata for the collections of the National Libraries of the Czech Republic, France, Poland and Slovenia is accessible via the protocol, as is that for the collection of monographs that will be OCR-ed by the National Széchényi Library in Hungary. Others are accessible via the

SRU protocol or via collection descriptions, but most collections concerned are currently only accessible through the website of the respective national library.

This result is not altogether surprising, as WP 1 was specifically designed to provide better access for collection with little or no metadata, but it will have to be taken into account when planning for Task 1.5.

Table 5: Overview of OAI-PMH accessibility of material to be OCR-ed

National Library	Material	Accessible via OAI-PMH
Austria	newspapers	not accessible at all
	governmental material	
Czech Republic	books	yes
	newspapers	
Estonia	newspapers	yes
	journals	not accessible at all
France	books	yes
	periodicals	
Hungary	periodicals	not accessible at all
	newspapers	not accessible at all
	journals	not accessible at all
	books	not accessible at all
	monographs	yes
	pamphlets	not accessible at all
Iceland	journals	no (via SRU)
	newspapers	
Latvia	books	yes
	newspapers	not accessible at all
Lithuania	newspapers	not accessible at all
Norway	books	not accessible at all
	journals	
Poland	books	yes
	newspapers	

Slovak Republic	newspapers	n/a as digital images are in the process of being produced
	journals	
Slovenia	newspaper	yes
	journals	
	books	
Spain	newspapers	not accessible at all
	books	
Sweden	newspapers	not accessible at all
	journals	
	books	
	printed ephemera	

5. Summary / outlook

This document's object is to describe the results of TELplus WP 1 Task 1.1, the "Survey of availability of digitised images for OCR". It is the first of two scheduled for WP 1. The second, which is the object of Task 1.2, has a much larger scope than this first one. It will be focused on technical background and file formats.

While there was a combined questionnaire for the two surveys, deliverable 1.2 will have a broader input base. The CENL / EDLproject Digitisation Survey and most prominently the WP 1 workshop on good OCR practice hosted by BnF in January 2008 in Paris, will be as much part of the basis for D 1.2 as the combined questionnaire and this current surveys results.

This deliverable, D 1.1, was designed to receive an idea of the material which would be OCR-ed by the WP 1 partners. It was also intended to give an idea of existing OCR experience and of how the original non-OCR-ed collections are accessible to The European Library. This last is an important first step towards planning for Task 1.5, "Provision of access to newly OCR-ed material through The European Library", which will formally commence in March 2008.

The knowledge of expected material and of OCR experience among the partners will fill a number of purposes. Firstly, it will provide a background on which to plan for the Paris workshop. It will enable the speakers to prepare specifically for the needs and material of WP 1 partners.

Secondly, together with the workshop, it will provide the basis for Task 1.3, "Identification of concrete materials for OCR, OCR specifications, implementation plans and tenders", and through it for "Carrying out OCR" in Task 1.4.

Thus, this document lays the groundwork for the whole OCR workpackage of TELplus and will be part of the foundation of its success.

In January a new FP7 project will start. This project, which is called IMPACT (IMproving ACcess to Text; <http://www.impact-project.eu/>), is part of the European Digital Library. “IMPACT as a network of centres of competence brings together fifteen national and regional libraries, research institutions and commercial suppliers. It aims to significantly improve access to historical text and to take away the barriers that stand in the way of the mass digitization of the European cultural heritage. „Within IMPACT the OCR challenges we are facing will be a mayor focus.. The National Library of The Netherlands (KB), the leader of the IMPACT project, through its The European Library staff is also a partner of WP 1, but does not provide content. With the KB and the National Libraries of Austria and France, we have three WP 1 partners in the IMPACT project, so that we can expect a lively exchange of ideas. Not only will WP 1 gain from IMPACT’s work, but we also expect that the information gain through the two surveys of WP 1, IMPACT will also gain from TELplus.