IMPACT - D-EXT2 PILOT REPORTS

Dissemination level: PU (Public)

CONTENTS

- Introduction
- Pilot reports
  - Functional Extension Parser (UIBK)
  - Error profiler and Post Correction Module – LMU (BSB, UA, KB)
  - Post correction tools – KB (LMU, IBM, INL, KB, University of Rouen)
  - ABBYY Finereader10 vs. Tesseract – PSNC
  - Dewarping pilot report – USAL
  - Retrieval Application - INL

Introduction

IMPACT has tested tools in productive environments in the last half year of the projects, e.g. with pilots of the CONCERT tools at the KB, BSB and the BL. Within the extension we have encouraged the uptake of mature tools and resources by running a number of pilots, involving more tools as well as more content holders. Each pilot has been described extensively in a pilot report.
Pilots

Functional Extension Parser – UIBK (DNB)
UIBK carried out two pilots during the extension of the IMPACT project. The two main objectives of these pilots were to:

- Integrate the FEP into the digitisation workflow of the eBooks on Demand EOD Network.
- Use the FEP platform for a new application area, e.g. the automated extraction of metadata from title pages of doctoral thesis for the German National Library (DNB).

Both pilots were carried out successfully and the objectives were achieved.

Error profiler and Post Correction Module – LMU (BSB, UA, KB)
In February and March of 2012 the LMU IMPACT team conducted a series of user experiments together the three IMPACT partners of Bavarian State Library (BSB), Dutch National Library (KB) and the University of Alicante (UA). The goal of these experiments was to evaluate the capabilities of the CIS postcorrection system in comparison to other correction systems, to see how much the error profiles and batch-correction features introduced by the tool can improve the correction process. The pilot was also used to gather information about the usability of the tool and possible ways of improving it.

Post correction tools – KB (LMU, IBM, INL, KB, University of Rouen)
The KB has digitised more than 2 million pages of 10,000 books from 1780 – 1800 for the project Early Dutch Books Online (EDBO). Due to the disappointing OCR quality, the Project Board decided to have students manually correct as much text as possible. Next to this manual correction, a pilot was done together with IMPACT to look at the various possibilities to enhance the OCR with the help of the following tools:

- CIS Postcorrection system with error profiling (TR5)
- Re-OCR-ing with:
  - ABBYY FRE 10
  - ABBYY FRE 10 and Dutch historical lexicon
  - IBM’s Adaptive OCR
- Alto Edit, a tool developed at the KB
- PlaIR platform from the University of Rouen (an improved version of the Trove newspaper tool)

The pilot also tested the CONCERT system made by IBM, but due to the setup of the pilot the test did not do justice to CONCERT and the results were not representative of the tool. Thus, the choice was made to not use the CONCERT output in this evaluation, but instead look at the results when using IBM’s Adaptive OCR compared to ABBYY 10.

The purpose of the experiment was to get a better view on the various OCR correction tools that are available, the amount of effort needed from a library and the volunteers to use such tools, and the quality of the results that can be obtained.

1 http://www.earlydutchbooksonline.nl/
ABBYY Finereader10 vs. Tesseract – PSNC
PSNC undertook a pilot to compare the OCR accuracy of two well known OCR engines: Tesseract 3.0.1 and FineReader10 Corporate Edition. The comparison is based on Polish historical printed documents and ground-truth produced within the scope of the IMPACT project.

Dewarping pilot report – USAL
This report summarises the results of the Dewarping Pilot which was carried out in the scope of the 2012 IMPACT extension. The main goals of this pilot were to investigate the potential of dewarping for image enhancement in general and as a pre-processing step in OCR workflows in particular, based on the tools developed in IMPACT. A secondary (but very important goal) was to investigate different evaluation methodologies for dewarping evaluation.

Retrieval Application - INL
INL focused on several aspects during the extension:
- Lexicon deployment in text recognition
  - Productization of external dictionary interface. For CCS GmbH, the historical Dutch OCR lexicon has been packaged with enhanced documentation and an updated version of the INL external dictionary interface implementation.
  - OCR with Named Entity lexica. A report on the contribution of the NE lexica to text recognition.
  - Using morphology in OCR. Test and evaluation of the module developed for deploying finite state morphology in OCR during the project.
- Using the INL Retrieval application developed in IMPACT, INL validated improved access. The emphasis was on scalability and deployment in real-life situations, with the addition / enhancement of several functionalities. For the specific languages, the following work was done:
  - Dutch: the Early Dutch Books Online (EDBO) collection, consisting of about 2 million pages from 10,000 books, has been indexed and deployed in the application to put the scalability to the test. This set has been the main test set for the optimization of the application.
  - Polish: pilot based on the Nowe Ateny encyclopedia, involving the use of the Polish IR lexicon.
  - Spanish: pilot based on data from the Biblioteca Virtual Miguel de Cervantes.
# Functional Extension Parser – Pilots

## Table of Contents

1. **EXECUTIVE SUMMARY**  
2. **EOD INTEGRATION**  
   2.1. Introduction: The EOD Network  
   2.2. The EOD Workflow  
   2.3. FEP Integration into EOD Workflow  
3. **DNB TITLE PAGE ANALYSIS**  
   3.1. Problem Statement  
   3.2. Metadata to be extracted  
   3.3. Technical Solution  
   3.4. Detection procedure  
   3.5. Ground Truth  
   3.6. Evaluation  
   3.7. Discussion of results  

IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.
1. Executive Summary

UIBK carried out two pilots during the extension of the IMPACT project. The two main objectives of these pilots were:

a) Integrate the FEP into the digitisation workflow of the eBooks on Demand EOD Network.

b) Use the FEP platform for a new application area, e.g. the automated extraction of metadata from title pages of doctoral thesis for the German National Library (DNB).

Both pilots were carried out successfully and the objectives were achieved.

The implementation of FEP within the workflow of the EOD Network was demonstrated at the conference Semantic technologies in libraries: from text to structure – from words to meaning hosted by the Research Library in Olomouc, Czech Republic on 10/11th May 2012. Nearly 10 EOD libraries declared their interest in the application and a follow up workshop will be held at Regensburg on the 11th of June in order to train the operators of the libraries. The pilot will be continued after the IMPACT project.

The second pilot was carried out with the DNB. Final results are considered to be convincing therefore a two days workshop will be held on 12/13th June in order to discuss the further potential of FEP for the production workflow of the DNB. Afterwards it is planned that an internal feasibility study will be issued by DNB in order to specify the implementation requirements.

A workshop was also held with the Austrian National Library on 16th March 2012. The decision on the further usage of FEP within the Google Books Project (Austrian Books Online) will be taken until September 2012. Another workshop is planned for August or September with the Finish National Library at the Mikkeli digitisation centre.
2. EOD Integration

2.1. Introduction: The EOD Network

European libraries are hosting millions of books published from 1500 to 1900. Due to their age and value, they are often only accessible to users actually present at these libraries.

With the EOD service these books become available to everybody - just a few mouse clicks away. Users are able to order eBooks via library catalogues; the libraries then digitise the requested items and send it to the user via the EOD service network. As of 2009, these books can also be ordered as a reprint in the form of a "real" trade paper book in addition to the digital version. EOD was first launched in October 2006 with 13 libraries from 8 European countries. By July 2008, a self-sustained network was established. Since then, other libraries from other countries have also been continuously joining the network so that currently more than 30 libraries are offering this digitisation on demand workflow.

2.2. The EOD Workflow

Briefly, EOD functions as follows. The starting point is either the online catalogue of the participating library or the EOD Search interface (a central index of most library catalogues). The EOD button is visible on the catalogue entry for all items available for digitisation. A user interested in obtaining a copy of a book simply clicks on the EOD button to initiate the process (see Figure 1).
Next, the user fills out the order form, where they can choose the type of delivery that they require. This may take the form of a download or a data carrier such as DVD. After this, the user is invited to follow the process of the order on their personal tracking page.

The library receives the order in real time, then scans the requested book and transfers the scanned images via FTP to the central server located at Innsbruck University Library. Each library manages and processes its own orders using a central database, the Order Data Manager, accessible via a web browser.

The workflow for the creation of the digitised book is also tracked via the Order Data Manager (see below). After completing the payment process, which supports online credit card payment, the user downloads the PDF from his personal tracking page. After a certain period of time, the library adds the digitised book to its digital library or repository and thus makes it available to the general public.

The Order Data Manager includes another core component: the Digital Object Generator, which is used to create eBooks by applying OCR and generating the PDF file (see Figure 4).
The Digital Object Generator supports OCR of all common typefaces from the 18th to the 20th centuries, including Roman and Gothic fonts as well as Cyrillic, Hebrew and Greek scripts, all of which are important when digitising a wide range of historic texts. It is also possible to generate a cover automatically, including the relevant metadata and displaying the logo of the delivering library. The creation of output files such as PDF, RTF and OCR XML, as well as the generation of the streaming link for downloading, is rendered during the process. The PDF eBook delivered to the user consists of two layers: the first layer containing the scanned images and the second layer containing the automatically recognised text.

2.3. FEP Integration into EOD Workflow

The Integration of the FEP System into the EOD workflow was, from the technical point of view, quite simple and straightforward. The web services which were developed within IMPACT as part of OC5 Impact Interoperability Framework were used as interface between the EOD system and the FEP. The EOD system calls the provided web services at the end of the digitisation process. Moreover EOD has to ensure that the output files of the EOD workflow system are accessible for the FEP system over the WEB (FTP, HTTP). The integration was planned and enforced in the following three phases:

Phase 1 (M1-M2): The main focus within the first phase was to extend the FEP web application with user handling and to improve the existing application mainly in terms of usability. Concerning the user handling we decided to keep it relatively simple. Each
partner library of the EOD network gets a username and a password for the FEP web application and there can only be one active session for each user. When the user logs into the web application, only books which were uploaded by the user’s institution are visible and ready for correction and export.

**Phase 2 (M3-M4):** The second phase was coined by testing the integration. The EOD team was able to use the provided web services and to load several ordered books into the FEP system. Some minor technical problems were identified and fixed within this phase. Another important test case was to get some “real user” feedback. Non-technical staff of the EOD team was asked to do extensive testing of the web application and to provide feedback about flaws or desired improvements and features. The most serious complaints and relevant feature requests were then implemented in order to provide real EOD network users an optimised FEP environment.

**Phase 3 (M5-M6):** The integration and the FEP system were presented to the members of the EOD Network at the EOD Workshop in Olomouc on May 11/12th. It was planned to select three voluntary partners, but actually many more libraries are willing to test and use the system. Therefore a separate workshop will be organised in June at the University Library Regensburg. We are expecting some useful feedback from these partners and will use their critics to stabilise and to improve the application. This phase will last until the end of June. The FEP service will then be supported by the team for 2012. During that time further decisions concerning the support and further financing of the service will be taken.
3. DNB Title Page Analysis

3.1. Problem Statement

The German National Library (DNB) still receives a large number of doctoral thesis (dissertations) in printed format. These dissertations are currently manually indexed by library staff which is a time consuming process. In order to support the librarians, a pilot was developed for using the FEP platform to automate this process. Metadata such as the title, publication date, the creator or the name of the university are extracted and (partly) normalised. The main question of this pilot was to find out if the automated process will effectively reduce the work load of the indexers.

3.2. Metadata to be extracted

DNB provided a list with metadata fields to be extracted automatically from the scanned images.

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Example</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author / Creator</td>
<td>Müller, Peter</td>
<td>The name appears in general as first name – second name, but has to be normalised to second name, first name.</td>
</tr>
<tr>
<td>Title</td>
<td>Main title, sub title and title additions</td>
<td>If title and addition to the title can be distinguished by the application, separation occurs in the title field as follows: Title &lt;space colon space&gt; title additions, in the title field also the creator is added by name as written on the title page, also including any introductory phrases (i.e. &quot;presented by&quot;, “vorgelegt von”) using &lt;space slash space&gt;</td>
</tr>
<tr>
<td>Publication Year</td>
<td>2010</td>
<td>Generally the date which</td>
</tr>
</tbody>
</table>
occurs without any additional indication.
In rare cases the date is introduced by specific phrases: “Tag der Einreichung”, “eingereicht am”, “vorgelegt JJJJ von”

<table>
<thead>
<tr>
<th>Language Code</th>
<th>ger, eng</th>
<th>Language of the thesis. Mainly German and English, in rare cases French and Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Format</td>
<td>30 cm</td>
<td>Physical format</td>
</tr>
<tr>
<td></td>
<td>21 cm</td>
<td></td>
</tr>
<tr>
<td>University publication notice record</td>
<td>Consists of multiple fields separated by commas</td>
<td></td>
</tr>
<tr>
<td></td>
<td>city, higher education institution type, document type, promotion year</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Not for the exchange)</td>
<td></td>
</tr>
<tr>
<td>Example:</td>
<td>Braunschweig, Technische Univ., Diss, 2010 (Only limited to the exchange)</td>
<td></td>
</tr>
<tr>
<td>Example:</td>
<td>Bochum, Univ., Habil-Schr., 2009</td>
<td></td>
</tr>
<tr>
<td></td>
<td>City and Higher Education Institution type taken from a given list in a standardised form.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Then often followed by the standard specification for the document type (Habil-Schr. or Diss.)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Then often followed by the promotion year, not to be mismatched with the year of publication (but promotion year is in 80% identical to publication year)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>If the title page is marked with a stamp exchange modalities follow in parentheses: a) (not for the exchange) b) (Only limited to the exchange) which are standardised information (only possible variants are a and b.).</td>
<td></td>
</tr>
</tbody>
</table>
3.3. Technical Solution

For the development of the prototype we applied the FEP framework. At first the ABBYY FineReader OCR SDK 10 was used to create OCR XML files for the 1000 title pages provided by the DNB. Then those title pages were inserted into the FEP database using the FEP ingest module. The first 300 pages were used to train the rule set and the remaining 700 pages were used for evaluation purposes.

The rules were formulated in JESS and applied with the aid of the FEP JESS module to the input title pages.

In addition dictionaries were used, therefore the FEP JESS module was expanded by the appropriate functionality. The dictionary search itself was implemented using the string library which was developed within the IMPACT project by the LMU in TR5 Post-correction. This software library was integrated through a specific interface into the FEP JESS module. The data for the dictionaries itself were partly directly taken from catalogues of the DNB. E.g. for the preparation of a comprehensive dictionary for authors/creators the full set of first names was read from the DNB database via an SRU interface.

3.4. Detection procedure and rules

At first every word of the title page is tagged using different dictionaries, e.g. for place names, first names, academic grades, university names and labels for the different fields.

Then rules are applied to identify the single structure types. Structure types that can be identified easily are treated first, like the document type (“Dissertation” or “Habilitation”), up to the difficult ones like the name of the university or the title of the dissertation.

In a next step already identified structure types are then used to identify the other types. The order and used structure types are so far:

1. VorgelegtVonType - a label introducing the creator
2. DocumentType – Dissertation or Habilitation
3. PromotionType and PublishType – promotion date field and publication date field
4. CreatorType – contains the creator/author
5. DepartmentPhraseType – contains faculty, institute, chairman and city of the university
6. GradePhraseType – contains the reached academic grade, spoken phrase and abbreviations
7. Revisiting DocumentType and DepartmentPhraseType – at this point those fields can be extended/created and checked
8. UniType – Contains the name of the higher education institution (university)
9. Junk - after every structure type except title was found, the rest of the text should consist of the title + some junk only, so by removing junk it is very likely that the rest contains the title
10. TitleType – contains title and subtitle

Once these processing steps are finished we receive the information of the title page. The next step is therefore to normalise some of the data. In most cases this is rather simple except for the name of the university due to many variations in spelling.

### 3.5. Ground Truth

To evaluate the quality of our automated annotation a ground truth (GT) data set was set up in collaboration with a service provider. The GT contains additional metadata fields to check the analysed introductory phrases for each specified metadata field. The GT was generated semi-automatically: The service provider created the segmentation for the fields using the Aletheia tool, which was developed by USAL in OC3 Metrics. By aligning that segmentation the corresponding text was extracted from the OCR XML files. This has the advantage that the OCR errors do not affect the evaluation of the results. On the other hand it has to be noted that with this procedure also OCR errors (which are very low in these modern texts) may occur in the GT.

### 3.6. Evaluation

The results of the analysis are matched against the semi-automatically created GT using f-Measure evaluation. Total number of analysed title pages was 997, three given title pages could not be downloaded from the DNB web site and two title pages were excluded from the GT because of some errors in the GT.

<table>
<thead>
<tr>
<th>Structure type</th>
<th>Precision</th>
<th>Recall</th>
<th>f-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title complete¹</td>
<td>87.03</td>
<td>82.17</td>
<td>84.53</td>
</tr>
<tr>
<td>Title contains²</td>
<td>100.00</td>
<td>94.40</td>
<td>97.12</td>
</tr>
<tr>
<td>CreatorType</td>
<td>98.51</td>
<td>90.17</td>
<td>94.16</td>
</tr>
<tr>
<td>DocumentType</td>
<td>97.28</td>
<td>99.65</td>
<td>98.45</td>
</tr>
</tbody>
</table>
Evaluation results for the test set for each metadata field (structure-type):

<table>
<thead>
<tr>
<th>Structure type</th>
<th>Precision</th>
<th>Recall</th>
<th>f-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title complete¹</td>
<td>85.02</td>
<td>78.00</td>
<td>81.36</td>
</tr>
<tr>
<td>Title contains²</td>
<td>99.27</td>
<td>92.83</td>
<td>95.94</td>
</tr>
<tr>
<td>CreatorType</td>
<td>98.33</td>
<td>90.17</td>
<td>94.13</td>
</tr>
<tr>
<td>DocumentType</td>
<td>97.16</td>
<td>99.89</td>
<td>98.51</td>
</tr>
<tr>
<td>PromotionType</td>
<td>88.33</td>
<td>99.87</td>
<td>93.75</td>
</tr>
<tr>
<td>PublishType</td>
<td>88.12</td>
<td>99.87</td>
<td>93.75</td>
</tr>
<tr>
<td>UniType</td>
<td>61.09</td>
<td>71.23</td>
<td>65.77</td>
</tr>
<tr>
<td>Overall</td>
<td>90.49</td>
<td>93.33</td>
<td>91.89</td>
</tr>
</tbody>
</table>

¹) GT title and analysed title are equal  ²) GT title is fully contained in the analysed title

### 3.7. Discussion of results

The accuracy of every structure type field could be improved by applying further rules, but it has to be mentioned that every added rule raises the complexity and has to be tested for their compatibility with the existing rule set. Nevertheless the current approach takes only textual data whereas layout data (e.g. coordinates of blocks and words) would be available as well. A separate rule set could utilise this information and results could be compared via a voting system.

We could not apply machine learning for the generation of the rules set since the GT was provided rather late. It is to expect that results could be improved by machine learning algorithms. The GT also could be used for the creation of richer and more appropriate dictionaries. Especially in the case of the identification of the higher education institution (university) this would improve the results significantly.