Automatic analysis of international newspapers, periodicals and press dossiers – A case study from Germany

Author/presenter: Dr. Juergen Warmbrunn, juergen.warmbrunn@herder-institut.de

Copresenter: Rolf Rasche, rasche@imageware

1 Abstract:

The Herder Institute possesses the largest collection of newspaper clippings focusing on Eastern Europe in the German-speaking countries. More than 6 million newspaper clippings from mostly the 1950s to 1990s are a treasure for research on the "socialist experiment" in East Central Europe. The collection is arranged by themes, persons, and locations. Newspapers from about 15 East and West European countries and in more than 10 languages were regularly analyzed and stored.

The press archives collection is now in the process of being digitized. A major question in this context is how to make the content easily accessible in a convenient way (i.e. in accordance with the conditions reigning in the "digitized world"). The very strict German copyright laws are an additional hurdle in the process of tapping the full potential offered by OCR.

Therefore some experienced software houses were asked to develop a flexible software tool able to generate meta-data from various source materials and to make the digitized content easily accessible.

The basic problems encountered before and beyond the mere process of digitization turned out to be

- the different sources (newspapers, journals, periodicals, newspaper clippings/dossiers),
- different paper types and qualities,
- different paper sizes, individual graphic designs and compositions of each individual newspaper or journal,
- semantic aspects (content): texts in one language, bi- or multilingual texts,
- different typefaces as well as
- different directions of reading and writing (e.g. Hebrew texts).

The software tool needed for this complex task should therefore cover the following key functions:

- preparation of the images to fit a specific internal standard (scans, e-paper)
- analysis and automatic marking of the different elements such as headlines, text, photos with text underlines, charts, diagrams,
- possibility of data output in different formats like PDF/XML, Mets, Alto, and others.

This paper will give a first insight into these difficulties as well as the results achieved.

2 About Herder and ImageWare

The **Herder Institute** was founded in 1950 in the old university city of Marburg in the state of Hessen. It served as the institutional substructure for the learned society Johann Gottfried Herder Research Council – an association of academics from various disciplines (mostly the humanities, but also law, economics, social sciences etc.) who had either their biographical roots or a strong research interest in the former East German territories lost after World War II.

Since 1952 the Herder Institute has maintained a newspaper collection and a newspaper clippings archives analyzing the East Central European press (with a special emphasis on minority newspapers) as well as comparable West European newspapers. Mostly because of a lack of physical storage space microfilming of newspaper cuttings was begun in 1995 and in the beginning of 1997 retrospective instead of prospective analysis was increasingly aimed at. In 1998 budget cuts made a reduction of newspaper subscriptions necessary. Following a recommendation by the institute's academic advisory board the formerly largely autonomous Newspaper Archives lost its status as a department of its own and was integrated into the research library. Following a reappraisal of the collection's worth and relevance the research library in 2009 began to develop a concept for the future digitization of the archives and to acquire additional existing newspaper and newspaper clippings archives, which complemented the existing ones either thematically or geographically. In 2010/11 a first data base version of excerpts from the newspaper clippings archives was made available online and since 2013 the persons covered by the newspaper clippings archives are integrated into the Herder-Institute's central registry of persons, a joint effort to make all information on personae of relevance to the history and culture of East Central Europe held at the Herder-Institute accessible by using the German National Authority File.

ImageWare is a software house for academic libraries. It specializes in workflows with interfaces to common library systems like OCLC and ExLibris. Its main emphasis is on solutions for ILL, direct delivery services, mass digitization and automatic formal cataloguing for journals, newspapers and monographs as well as the presentation of copyrighted materials in ways conforming to the locally applicable copyright legislations**.**

3 Introduction

The project aims at making the Herder-Institute's newspaper clippings archives on East Central Europe accessible in the digital world by offering possibilities of research for academia through a quality-controlled digitization of the cuttings and a formal analysis of each clipping.

Due to the sheer volume of around 5.3 million newspaper clippings efficient procedures for mass digitization must be put in place. Each clipping has a unique character: not only due to its composition (format, paper quality etc.) but also because of its content. As a consequence it is not feasible to digitize and analyze several newspaper cuttings in one step. Analyzing 5.3 million newspaper clippings thus probably equals analyzing 5.3 million monographs.

Classical procedures for formal and content indexing can thus be ruled out. Since the individual newspaper clippings have as yet not found their way into the OPAC no metadata can be used as a basis or be put together with the digitized images.

At an early stage of the project automatic indexing with simply OCR technology seemed a possible solution in spite of the rather limited depth of content analysis this would necessarily have involved. It became quite clear, however, that this approach had to be abandoned for both technical and copyright reasons.

The technical objection was linked to the fact that in the case of newspaper clippings the relevance of a term and/or a name stems from its use within certain article segment, e.g. the headline, subtitle or the text of the article. Simple OCR cannot provide this additional information, all terms/names taken from a purely OCR generated text would receive the same relevance ranking or be ranked according to their frequency. A qualitative ranking is thus only possible when the occurrence of the search term in different article segments is properly taken into account.

A further technical objection was linked to the necessary consideration of holdings of the cooperation partners of the Herder Institute. The institute cooperates among others with the Foundation Martin Opitz Library in Herne and the Bavarian State Library. A long-term aim of the project is thus to make holdings of these and other cooperation partners available, taking into account possible duplicates or complementary holdings. To avoid double work a solution had to be found that makes duplicate control possible in spite of the evident lack of standard numbers (ISBN/ISSN/URN/DOI etc.), i.e. through the analysis of faultless indexed titles in conjunction with formal data like newspaper title, date/issue number, etc.

The copyright related objection stemmed mostly from an increasing feeling of uncertainty regarding the rights of authors and publishers that have to be taken into account. Whereas only a few years ago copyright questions seemed to be of little relevance as far as collections of newspaper cuttings were concerned the attendance of several legal and copyright

seminars soon showed the Herder-Institute's staff in charge of the project that copyright matters had an immediate significance when planning to make the holdings accessible in the digital world. The rights of authors and publishers are increasingly enforced and become the basis of law suits. Even huge enterprises like Google (cf. the legal dispute Google vs. Collecting Society Word) and large academic libraries (cf. the legal dispute Open University Hagen vs. Kröner Publishers), which serve research and academic teaching, are affected by this prevalent tendency.

Since the Herder Institute aims at a sustainable solution that can be maintained over years to come it cannot rely on legal "views", which differ according to the standpoint of the party concerned, but must introduce a system that offers the highest possible legal certainty. It should also be able to react to changes in the jurisdiction concerned and to put immediately into practice individual agreements with publishers/copyright holders in order to always comply with the current legal requirements. Typical examples for this are the blocking of photographs, the release of all articles of a certain publishing company or access only to content up to the year 1965 from outside the premises of the Herder Institute.

Since no standard solution on the market was found that could master these objections and deal with the project specific huge volume of articles, the Herder Institute decided to cooperate with two companies in order to create a new overall custom-sized solution from their standard solutions. One of the companies concerned is the Gesellschaft fuer angewandte Informationstechnik (GfaI) [Society for Applied Information Technology] from Berlin, which developed the algorithms for the segmentation of the newspaper cuttings which in turn form the basis for the automatic formal indexing. The other is the ImageWare Company from Bonn, which integrates these algorithms into their workflow solution for mass digitization MyBib eDoc. MyBib eDoc was already used by the Herder Institute's partner institution Bavarian State Library as a workflow solution in the Google project for more than one million monographs as well as by the Foundation Martin Opitz Library in Herne. The volume of around 250.000 to 300.000 annually digitization orders successfully reached within the Google project equals about a third-fold of the standard volume (i.e. without additional third-party funding) of processed newspaper clippings aimed at by the Herder Institute.

4 The task

4.1 Material from the Press Archives

The archives cover two periods: around 1920 and from 1952 to 1999. It consists of ca. 5.3 clippings in around 16.300 files (around 1.6 shelf kilometers) as well as 640.000 microfiches and 160 microfilms. The clippings cover mostly Poland (40 %), Czechoslovakia (40 %) and the Baltic countries (15 %). 60 % are topic related, 33 % person related and 7 % geographically related.

The Press Archives moreover contains the following separate collections:

- a special collection on the Inter-War Period
- the newspaper clippings archives of the German Institute for Contemporary History of the GDR (later Institute for International Politics and Economics of the GDR) (1952-1990; USSR, South Eastern Europe)
- Private Archives on Contemporary History and Politics/Wolfram Gabriel (born in Breslau/Wrocław, Rheinstetten-Neuburgweier (800 filmes with 30.000 clippings; 1982-1994)
- Radio, press and television reports from Radio Free Europe (1957-1991); ca. 1.260 archival boxes (600 on Poland, 570 on Czechoslovakia and 90 on South Eastern Europe) on more than 115 shelf meters

4.2 Demands regarding the scanning

The clippings are found in three different types of dossiers (persons, places, topics). Within the dossiers no clipping corresponds to the previous one, each clipping is unique due to its different paper quality, language, preparation (folded once or several times, stapled) and format. Digitization by means of a sheet feed flatbed scanner is out of the question since each clipping has to be handled and put on the scanner separately. As a consequence only A3 or A2 flatbed or overhead scanners can be used.

Since the quality of the OCR depends on the quality of the scans it is imperative that the documents are put on correctly. Post-scanning functions which correct slants also tend to lead to properly-adjusted images but they also result in artificial changes which in turn have a negative influence on segmentation and OCR during the further workflow. In other words: the better the scan quality at the beginning, the less necessity to work it over later and/or the less possibility of mistakes during the further process.

In the course of the tests the following further aspects became apparent:

1. 300 dpi are sufficient for the digitization; as a principle all scans should be made in color.
2. For the different paper types from different countries and eras specific scan profiles should be developed, taking into the account the respective specific characteristics. This makes it superfluous to make adjustments for every scan. The operator can perform the scan/rescan according to the document's relevant scan profiles.
3. Since many clippings are not rectangles but polygons the classical removal of black margins can unfortunately not be applied.fdc

Substantial information like e.g. dates, titles was often applied only in hand writing, partly in fading red or with pencil on the margins of the clippings. Since this information is highly

relevant for the legal assessment and the following indexing/processing the recording of this steering information should be as easy as possible.

4.3 Technical requirements

Because of the volume of work concerned the project's running time is scheduled to last for five to ten years. During this time the respective status of all orders as well as the status of the whole project must be comprehensible at any time. In order to achieve this objective two standardized workflows, one for the press dossiers and one of the digitized materials, have to be implemented. The workflow of a press dossier has to be documented from the selection in the archives over the scanning to the return to the articles, so that it is always apparent where each dossier can be found and it is guaranteed that each dossier is indeed returned to archives in the end. In the case of the digitized items the workflow begins with the scanning and continues with the indexing/segmentation at different workstations until the delivery to the presentation system. Further steps include quality control and the correction of mistakes. It also has to be taken into account that the indexing and the quality control require language skills. Therefore the newspaper clippings have to be delivered to the workstations where the operators have the language qualifications required. Since not all employees will be available for the whole duration of the project it is important that the training of new staff is easy and that the loss of staff members does not lead to limitations in the realization of the project or the loss of data, know how or preparatory work.

Given the limited personal capacities of the Herder Institute it must – in case additional third-party funding makes this possible – be possible to outsource parts of the scanning to external service providers. The work done by these service providers must, however, be recorded in exactly the same way as described above. This means that lists of the press dossiers which have been given to the external service providers must be provided, returns must be recorded and the produced digitized images must be linked accordingly. This is completely irrespective of the form of the press dossiers (paper or microform).

During long project runs not only the staff employed but unfortunately also the technical systems are liable to changes. For this reason a change of some or ALL system components must be possible AT ANY TIME. To guarantee this all images and all order data information including the descriptive and legal metadata are regularly exported in a well-documented machine-readable formal and saved independently from the systems. This should guarantee a certain independence from producers, which would enable the Herder Institute in the case of a system change to change from an old supplier to a new producer with special or conversion programs. This approach is in the beginning more demanding but in the light of the running time and the necessary independence from producers an indispensable prerequisite for a successful and sustainable project.

4.2 The legal environment

(Please note: The following remarks are based on our experience in the German legal situation. These can only in a very limited scope be applied to other countries. Given the international convergence of legal regulations, e.g. through European Union legislation, we can suppose, however, that similar regulations (in reduced or more severe form) exist also in other countries).

Decisive for the assessment of the permissibility of digitization is the definition of the term WORK. As long as only the complete newspaper/journal counts as a WORK, the digitization of an individual newspaper clipping, legally called “small part of a work”, represents no problem for libraries. Following the new possibilities for ordering single articles on the internet the legal term “work” has undergone a change, however, and a major problem for the digitization of newspaper cuttings has arisen. Currently no clear legal situation/judgment exist that would define whether newspaper cuttings are in the legal sense works of their own or not. According to a judgment by the Trade Law Court of the Canton of Zürich (HG110271-O dated 7 April 2014) we have to assume that in the next years also single newspaper clippings will be found to be original works. Due to the accessibility of individual newspaper clippings from the publishers and following the legal definition as an original work the relevant legal exception rule, which gave public libraries the right to digitize and make available excerpts from a work, can no longer be applied. To overstate the argument: As a consequence the Herder-Institute could then only digitize 10 % of the document, i.e. only the title of each newspaper cutting.

A newspaper clipping usually consists of both text and photograph, in some cases it is further enriched by a caricature or graphic. As a result its digitization involves **at least** two collecting societies: The Collecting Society Word for the text and the Collecting Society Picture for the photograph. Apart from the photographer’s copyright further rights can be concerned according to the depicted person or object. In the case of persons for example personality rights must be taken into account (the right to forget according to latest EU legislation), in the case of objects the copyright of the depicted object, of the architect, of the artist etc. As these rights are not valid “eternally”, there must be a possibility to make digitized photographs accessible or to block them according to more or less formal criteria.

A further copyright regulation which is relevant for indexing, whose effects are however often underestimated, concerns the right to duplicate. The right to duplicate is not only required for the analogue object but also for the electronic copy, in concrete terms for the working copy of the digitized image, which an OCR program needs to create the OCR. In short: without the right to duplicate there is no OCR.

It is not possible to foretell either the timeframe or the scope of future legal developments as far as the use of digitized images by libraries is concerned. Thus all eventualities that can

affect the workflow, the legal metadata or its presentation must at least be anticipated in order to be able to react efficiently when, for example, all photographs taken after the qualifying date X must be cut out or when all articles published after the qualifying date Y can be additionally indexed by means of OCR. The term “efficiently” means in this context that a legal change does not mean that all the 5.3 million cuttings must be re-examined in order to adapt them to the changed legal situation. In order to achieve this all relevant parts of a newspaper clipping (i.e. article texts, photos and others) must be segmented accordingly, the segments must be kept in machine readable form as a prerequisite for automatic carrying out of these processes. Irrespective of this the individual blocking of certain articles for example due to court judgments must be possible.

5 Pilot phase

5.1 Requirements and test series

As part of the pilot development a concentration on the article segmentation as a central and decisive component of the project took place. A test series took place that aimed at providing the basis for answering the following questions:

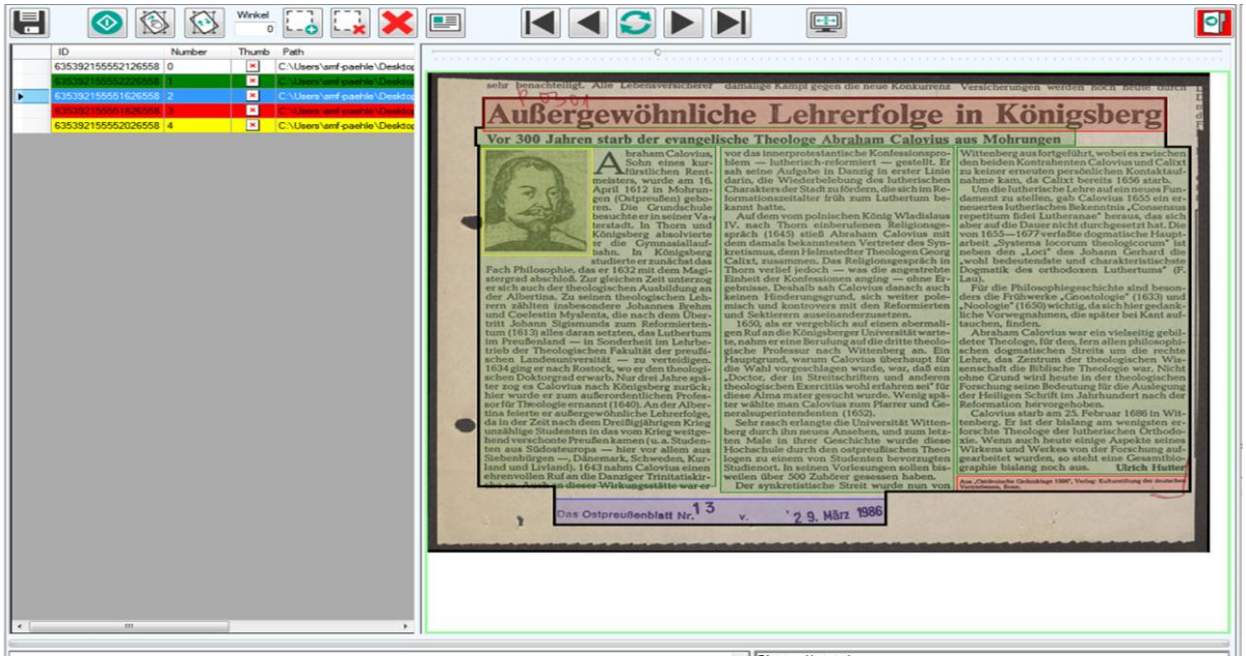
1. Does the segmentation actually work in the case of heterogeneous newspaper clippings?
2. To which extent are we faced with the need to manually correct the segmentation?
3. Can the processes take place as part of a batch process so that – given sufficient scan capacities – the articles can be segmented automatically?
4. Will segments that are needed in order to fulfill legal requirements (authorship of the article, of a photograph etc.) be recognized?

The test series was conducted on the basis of the dossiers for persons. As a consequence 34 files (person names from A-D) with about 9.000 clippings from different newspaper and in varying paper qualities and languages were processed. In parallel to the practical work conducted in the institute itself solutions for workflow and presentation from partners known to us were evaluated. The main criteria for the evaluation were whether individual or standard programs were involved and whether the system was reliable and constantly accessible. Also important were both the quantity of processed materials and the availability of interfaces that would guarantee the integration of the algorithms needed for the article segmentation.

5.2 Autosegmentation / clipping

A typical digitized document looks like this before segmentation:

In this case the result of the automatic segmentation process is as follows:



It can clearly be seen that the article text, the title and the photo/graphic were all clearly recognized. The title and the subtitle can be OCRed as segments of their own ("small parts of a work") in order to automate the indexing. Depending on the applicable right the article text will either be processed via OCR or the segment data will be stored in a way that makes it possible to process it automatically via OCR at a future date, once the rights needed have been obtained.

One aspect of the piloting, that had not received the attention required beforehand, was the lower part of the page, in which the title of the newspaper and the date are given (so-called recognition of the signet). This field is particularly important for the workflow. It must be reliably recognized and should be made readable via OCR. For hand-written dates an OCR was developed which delivers surprisingly good results if the handwritten numbers are comparatively readable.

During the piloting particular emphasis was put on an approach as pragmatic as possible. It was decided not to reduce the error rate, i.e. wrongly or not correctly recognized segments, that occurred in proportion to the quality of the document, through more and more refined algorithms. Instead more attention was invested into the usability with the aim of

1. First glance recognition of mistakes by the operator (in less than a second)
2. Manual correction of the mistake through a simple cooperation (< 10 seconds/segment).

Since every cutting has to be controlled as part of the quality control process this seems to be the most efficient way to achieve the results wanted.

As a result of the segmentation process all information needed for the further running of the processing is available:

1. Language for OCR and Quality control – Work place of staff
2. Newspaper title and date for checking of works out of print or without known author
3. Author for checking in the Common National Authority File (formerly Person Name Authority File).

On the basis of the heading and date the workflow system can conduct a duplicate check against the holdings of already segmented newspaper clippings.

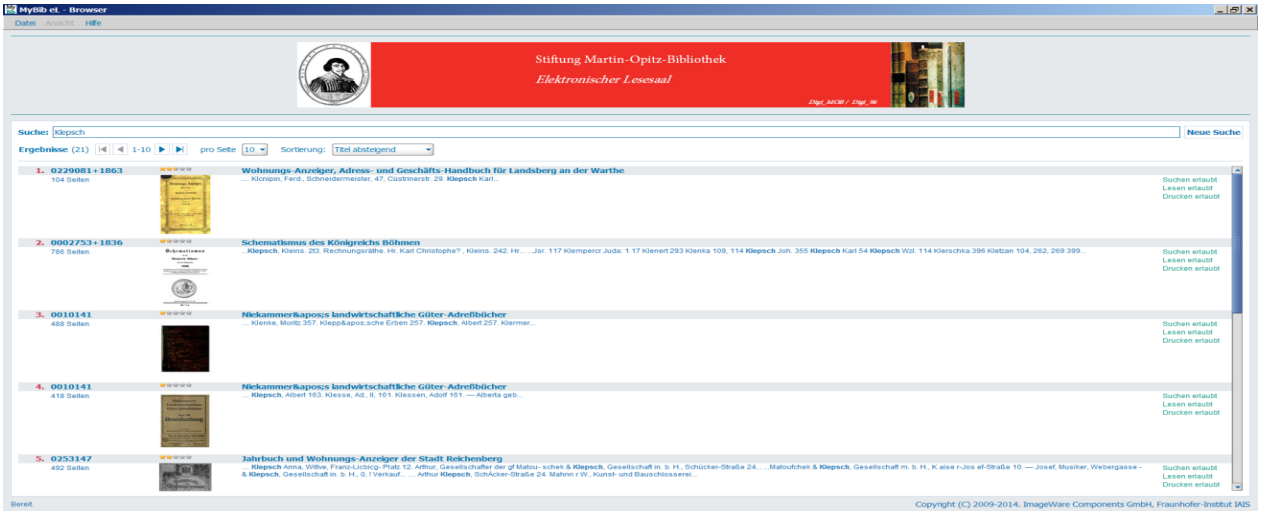
As part of a further development financed by our partner ImageWare an automatic page break function was developed. This, however, is for the time being problematic in terms of copyright law since changes to the work are only allowed with the author's consent.

This function will usually be of interest for forms of newspaper digitization when the respective article will be highlighted in yellow on the whole newspaper page shown as a thumbnail. This enables the user to identify directly where the article was situated in the original. For “folded” newspaper cuttings, which are larger than A 4, this page break function is useful in order to be able to show them on a single page.

The results are stored in a documented XML format, which can be analyzed and converted by easily available programs. In this way the central requirement of sustainability is fulfilled.

5.3 Pilot data base – Presentation

In order to evaluate the results the excerpts which had been indexed according to their individual segments via OCR were entered as full text in a SQL data bank (data base trial version) and a simple search mask added.



Search results are shown in this way:

Due to the legal problems that could not be foreseen by any of the project participants at the beginning of the initial project the data base search is only possible within the Herder Institute, no images are being shown and no searchable PDF is being produced.

5.4 Results

The segmentation works with heterogeneous newspaper clippings and supplies all relevant data for the legal and technical requirements as well as for the steering of the workflow. It functions automatically and can thus be run in batches. Each article can be viewed individually, the need for manual correction is small and can be further reduced through learning effects. The produced data can be further processed in XML format by additional programs. Thus the required components necessary for automatic indexing are available.

The outstanding requirements for the workflow like cooperative work and duplicate check can be solved by using the MyBib eDoc system, which has proved its merits in this regard through the use by among others the Google Project of the Bavarian State Library. Since 2006 it has also been successfully used for cooperative catalogue enrichment by several union catalogues and libraries with a data set of around 1.8 million titles.

As far as presentation systems are concerned we closely follow a project run by the Foundation Martin Opitz Library, which aims at presenting all digitized materials produced since 2000 in accordance with copyright requirements and which equals the legal complexities encountered by our project. To this purpose the Foundation Martin Opitz Library received project funding from the Commissioner for Culture and Media of the Federal Government with which it will create a presentation system for about 3.000 titles /500.000 pages according to the current legal stipulations. The holdings of the Foundation Martin Opitz Library comprise – analogue to our own holdings – copyright free titles, which can be accessible in the German Digital Library/Europeana, and copyrighted titles, which can be make available without OCR within the Foundation Martin Opitz Library.

6 Look ahead

Once the legal environment has been clarified we plan – subject to the availability of thir-party funding and the creation of the internal prerequisites – the following project realization:

1. Acquisition of the segmentation tool and going into production with the aim of indexing about 10.000 newspaper clippings this year. We aim at 100 scans/hour and 100 indexed images/hour. If we take into account the further steps of fetching the dossiers, scanning, indexing, and returning the files to the stocks this would mean 50 newspaper clippings/hour or 100.000 cuttings per annum per full time staff equivalent.

2. Introduction of a workflow solution in order to guarantee a sustainable, constant and scalable workflow over the next years. This would enable us to step up production if additional financial support makes a large investment into staff possible.

3. Parallel running of the present trial data base as a basis for the selection of a presentation system.

The main factors behind the selection and publication are purely legal requirement. The workflow system will allow us to determine which works are copyright free and can be presented in high quality on the internet. The remaining subset will be caught as images and partly indexed, i.e. with a zonal OCR but without an OCR of the article text. Unfortunately we will also only be able to show these images within the confines of the Herder-Institute. We look forward to presenting first results as part of a workshop in the Herder-Institute in the second half of 2015 when our current building project has been completed.