Submitted on: 8/16/2014

Making Historic Newspapers Available Online: Why, Where and How

Hans-Jörg Lieder
Department for Bibliographic Services, Staatsbibliothek zu Berlin – Preußischer Kulturbesitz, Berlin, Germany.
E-mail address: hans-joerg.lieder@sbb.spk-berlin.de

Abstract:

Within the Europeana Newspapers Project libraries and technology providers create and provide access to around 10 million digitised historical newspaper pages in the form of digital images and texts. The project also addresses issues related to all parts of the digitisation process and provides best practice recommendations for highly automated, effective workflows. In order to determine and create services and tools for the analysis of the digital newspaper text corpora libraries need to closely ally with users and developers and cooperate in practice. The necessary prerequisite for all future work, the availability of digital text corpora of reasonable quality, cannot be ensured exclusively by automated processing but rather requires a step-by-step approach including some degree of human intervention. Correcting texts in a joint effort at different places, i.e. local newspaper web pages and the European portals, requires sound technical solutions and non-technical agreements that can only be reached in close cooperation with user communities.

Keywords: Historic Newspapers, services, digital text corpora, OCR, layout analysis

Making Historic Newspapers Available Online: Why, Where and How

> Newspapers are the second hand of history. This hand, however, is usually not only of inferior metal to the other hands, it also seldom works properly. Arthur Schopenhauer

For librarians historic newspapers are double-edged objects: usually the originals were simply not meant to last. The quality of the crumbling, brittle and torn paper, fading ink, and less than perfect prints bear profound witness to the assumed ephemerality of the actual objects. As Schopenhauer said, “inferior metal”, indeed. Occasionally it seems that our librarian predecessors have contributed to these problems: missing issues, supplements, or

single pages, faulty bindings of the newspapers occasionally blocking out text, poor catalogue records – the list of lamentations concerning the physical objects seems endless.

Yet we experience an overwhelming interest on part of our users as soon as we make historic newspapers available online. Newspapers are, in a very real sense, "the second hand of history", the central place for almost all political and cultural discourse that manifests itself in the rapid succession of news, opinion, argument and counter-argument, propaganda, commentary, contextualisation etc. Thus newspapers allow a detailed analysis of the microstructure of history, of the constant to and fro that Schopenhauer could not perceive without notions of dysfunction and that eventually crystallises in parts in the shape of history books. Though various newspapers exist that are targeted at specific audiences, principally the thematic scope of newspapers is without limits. This makes them interesting for virtually all user groups. Researchers of all fields of scholarship but also the interested public will find relevant information, ranging from the description of historic events to local news and even the obituaries of family members.

In the light of such ubiquitous interest it is not surprising that a number of large initiatives provide access to vast quantities of newspapers. The "Chronicling America"[1] and the "TROVE Digitised Newspapers"[2] services are examples of such initiatives outside of Europe. In Europe, the continent with the richest newspaper tradition, we see a variety of projects – the ANNO portal of the Austrian National Library[3], the ZEFYS portal of the Berlin State Library[4], and the corresponding services of The British Library[5] and the National Library of Finland[6] are some examples of major newspaper web pages. Use conditions differ widely and range from the free presentation of public domain materials to access limited to registered library users and, in some cases, to access being liable to pay costs.

Europeana Newspapers

The Europeana Newspapers Project (ENP)[7] tries to support such European initiatives by defining procedures and workflows for the creation of content and by easing the digital resources' way to the European portals, namely Europeana[8] and The European Library (TEL).[9] The project's core mission is the provision of around 10 million digitised historic newspaper pages including full texts created by applying Optical Character Recognition (OCR), and 2 million pages including full texts that are segmented by article as a result of Optical Layout Recognition (OLR). Further to this the project also provides solutions or best practice recommendations concerning all steps involved in the digitisation of newspapers:[10]

- **Selection of newspapers to be digitised and determination of refinement methods** Quality prediction and quality evaluation software developed within the project help libraries in the a priori evaluation of expected results, thus enabling them to make

[1] http://chroniclingamerica.loc.gov
[2] https://trove.nla.gov.au/newspaper
[3] http://anno.onb.ac.at/
[4] http://zefys.staatsbibliothek-berlin.de/
[5] http://www.bl.uk/subjects/news-media
[6] http://www.nationallibrary.fi/services/digitaalisetkokoelmat/historiallinensanomalehtikirjasto17711890.html
[7] http://www.europeana-newspapers.eu/
[8] http://www.europeana.eu/
[9] http://www.theeuropeanlibrary.org/
[10] For more detailed information and access to the project's public documents and reports see the project's web pages.

informed, cost-effective selection decisions and determine a reasonable degree of refinement. This is done by creating some "ground truth" pages, i.e. digital pages that were manually corrected (text and page layout) following automated pre-production using FineReader. These near perfect results can be compared with results achieved by automated processes. The differences between both versions can then be finely analysed and evaluated with the help of sophisticated software that produces more accurate results than any human inspection could produce.

- **Creation of digital images and specifications for refinement**
  The project has developed software tools that enable libraries to easily transform and validate image files before subjecting them to OCR or OLR, thus ensuring that the highly automated refinement workflows are not impaired.

- **Large-scale, highly automated workflows for refinement**
  The major part of the OCR work undertaken in the project is performed by the Department for Digitisation and Digital Preservation of the University Innsbruck Library, an experienced service provider and project partner well equipped for mass digitisation and mass OCR projects. A smaller portion including OLR is undertaken by CCS [11], a commercial project partner with great experience in large-scale digitisation and conversion initiatives. Both service providers assure that workflows are highly automated and that improvements implemented during the project will be incorporated in their future services. On a smaller scale the project investigates the capabilities of Named Entity Recognition (NER) for selected languages.

- **Metadata best practice recommendations**
  Images, digital texts and mark-ups such as OLR and NER require metadata. The project investigates all necessary metadata and formulates best practice recommendations as the "European Newspapers METS[12]ALTO[13] Profile" (ENMAP). Of specific interest is the project's work on structural metadata that aims at enabling the classification of newspaper content as logical units, e.g. headline, caption, article, advertisement etc. It is expected that this work significantly contributes to a better understanding of the basic building blocks of newspapers and that such better understanding ultimately leads to more precise search options.

- **Transmission of results to Europeana, TEL, and the Union Catalogue of Serials**
  The full texts and digital images of newspapers are presented in TEL, metadata including links to the actual resources in local libraries are provided via Europeana and the German Union Catalogue of Serials (ZDB).[14] The transmission of large amounts of data to the European portals requires highly automated procedures with a

[11] http://www.content-conversion.com/
[12] Metadata Encoding and Transmission Standard, see: http://www.loc.gov/standards/mets/.
[13] Analysed Layout and Text Object, see: http://www.loc.gov/standards/alto/.
[14] The Union Catalogue of Serials / Zeitschriftendatenbank – ZDB (http://www.zeitschriftendatenbank.de/) is the world's largest dedicated database for serials. Currently ZDB contains around 1.7 million title level records and 13.6 million records describing local holdings in more than 4,400 libraries and other types of institutions, mainly in Germany and Austria. The integration of ENP data is an important step towards the internationalisation of the ZDB. The vision is clearly formulated by researcher Amélia Del Rosario Sanz Cabrerizo, professor at the Universidad Complutense de Madrid: "It could also be useful to have a centralised index of newspapers for every digital newspaper library which has aggregated data and material to Europeana" (from an interview conducted in July 2014; for the entire interview see: http://www.europeana-newspapers.eu/qa-with-newspaper-researchers-amelia-sanz-cabrerizo/).

minimum of human intervention. Workflows and specifications developed in the ENP assure that in future corresponding digitisation projects may fall back on well defined and established workflows. The data transfer between TEL, Europeana, and ZDB will continue after the end of the project.

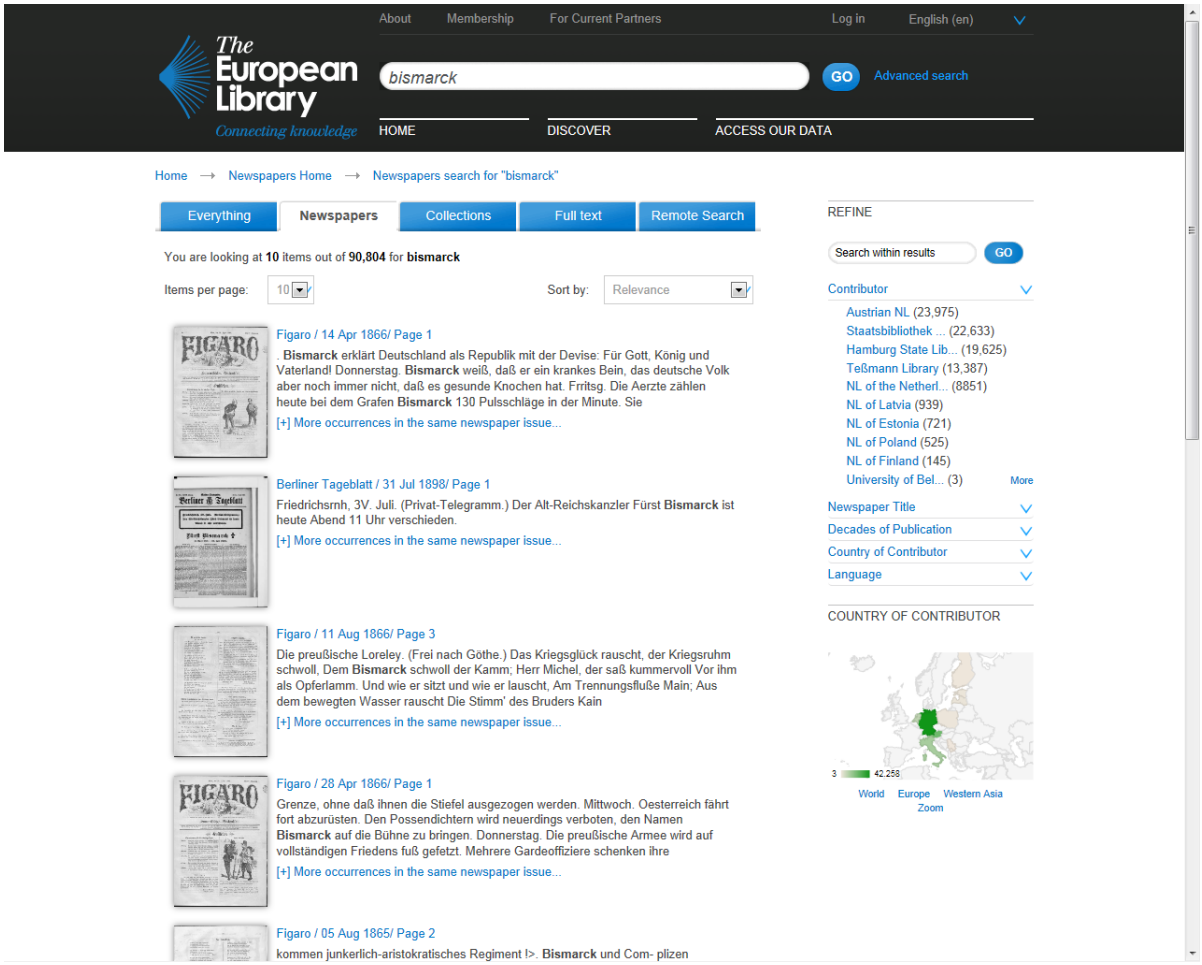

Fig. 1: Search results list in TEL

- **Presentation of results in the European portals**
  Digitised newspapers require a specific presentation environment. Prior to the ENP neither Europeana nor TEL provided functionalities to meet these requirements. Within the project search and browsing facilities were developed and added to TEL, which now provides access to the full extent of available newspaper data, texts and images, via a dedicated newspaper web page.[15] The "TEL newspaper browser" presents digital images of newspaper pages in combination with the texts created by OCR. It enables page-to-page navigation, a zoom function, highlighting of text blocks etc.
  Europeana provides a direct link from their search results page to TEL's newspaper browser. Users of Europeana can therefore access the actual data, seemingly without leaving the Europeana environment. This greatly contributes to a comfortable and seamless user experience in Europeana and, at the same time, avoids unnecessary data loads for Europeana's servers.

[15] http://www.theeuropeanlibrary.org/tel4/newspapers

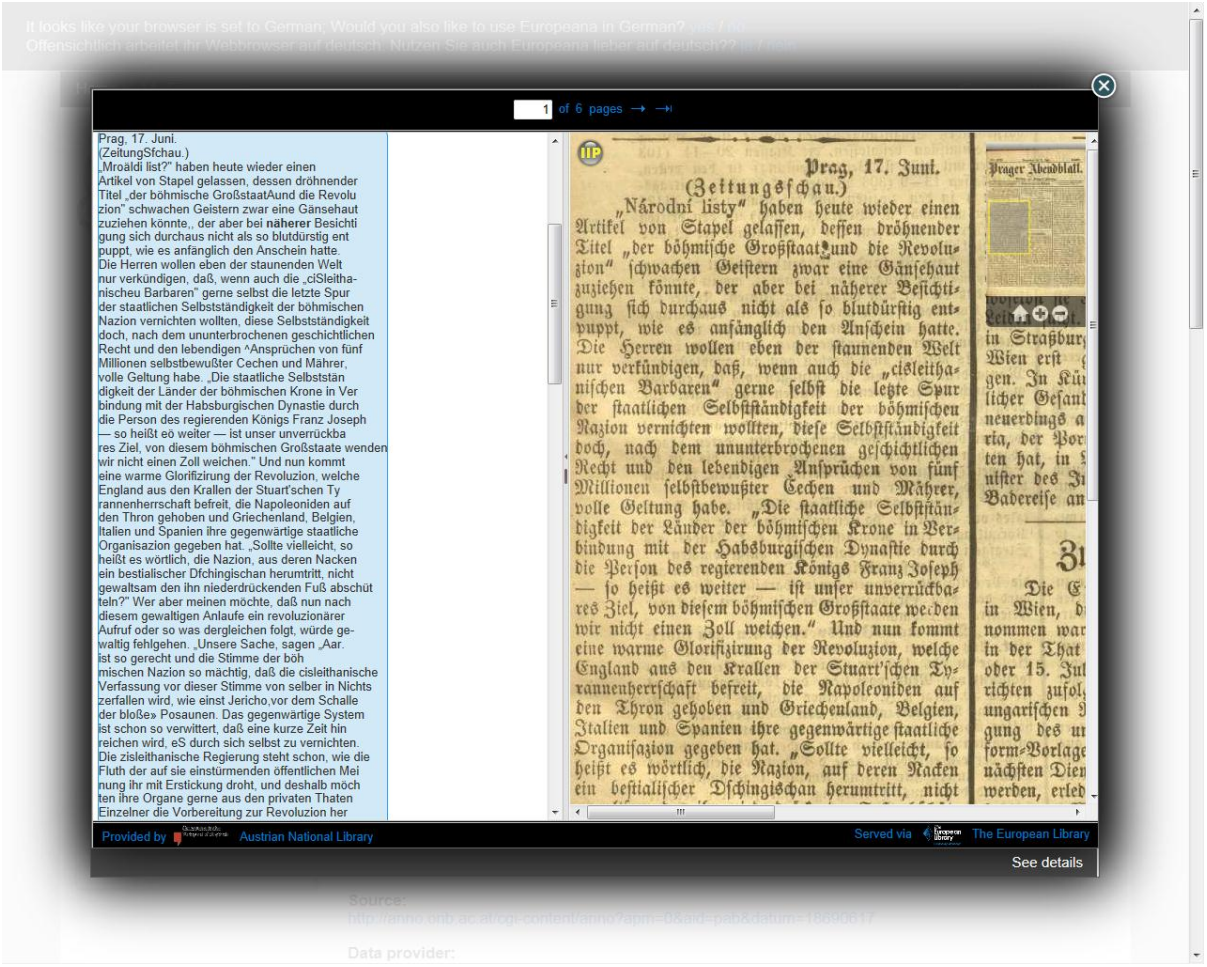

Fig. 2: View of a digitised newspaper page including automatically created text via the TEL newspaper browser as embedded in Europeana

An English language web page was added to the ZDB, enabling immediate access to digitised and digitally borne newspapers in European institutions.[16] Next to the metadata from project partners Europeana and ZDB will provide metadata from associated partners describing around 19 million additional newspaper pages.

[16] See: http://zdb-opac.de/LNG=EN/DIGI_NEWSPAPER. The ZDB will present the project data together with more data about digitised newspapers mainly in German and Austrian libraries, including resources that are not present in Europeana. Access to borne digital newspapers is also provided via this service that thus bridges the gap between freely available historic newspapers and modern resources. Note that the ZDB currently runs a standard OCLC-OPAC with functions and limitations typical for that product. Since the ZDB plays a major role in the organisation of digitisation projects in Germany and since German libraries have expressed their dissatisfaction with the current OPAC, it will be replaced by a more modern application that allows for many more functions. This work is funded by the German Research Foundation and will be finished by March next year. The "European Registry for Digital Newspapers" will be migrated to the new platform and benefit from the additional services independently of the ENP.

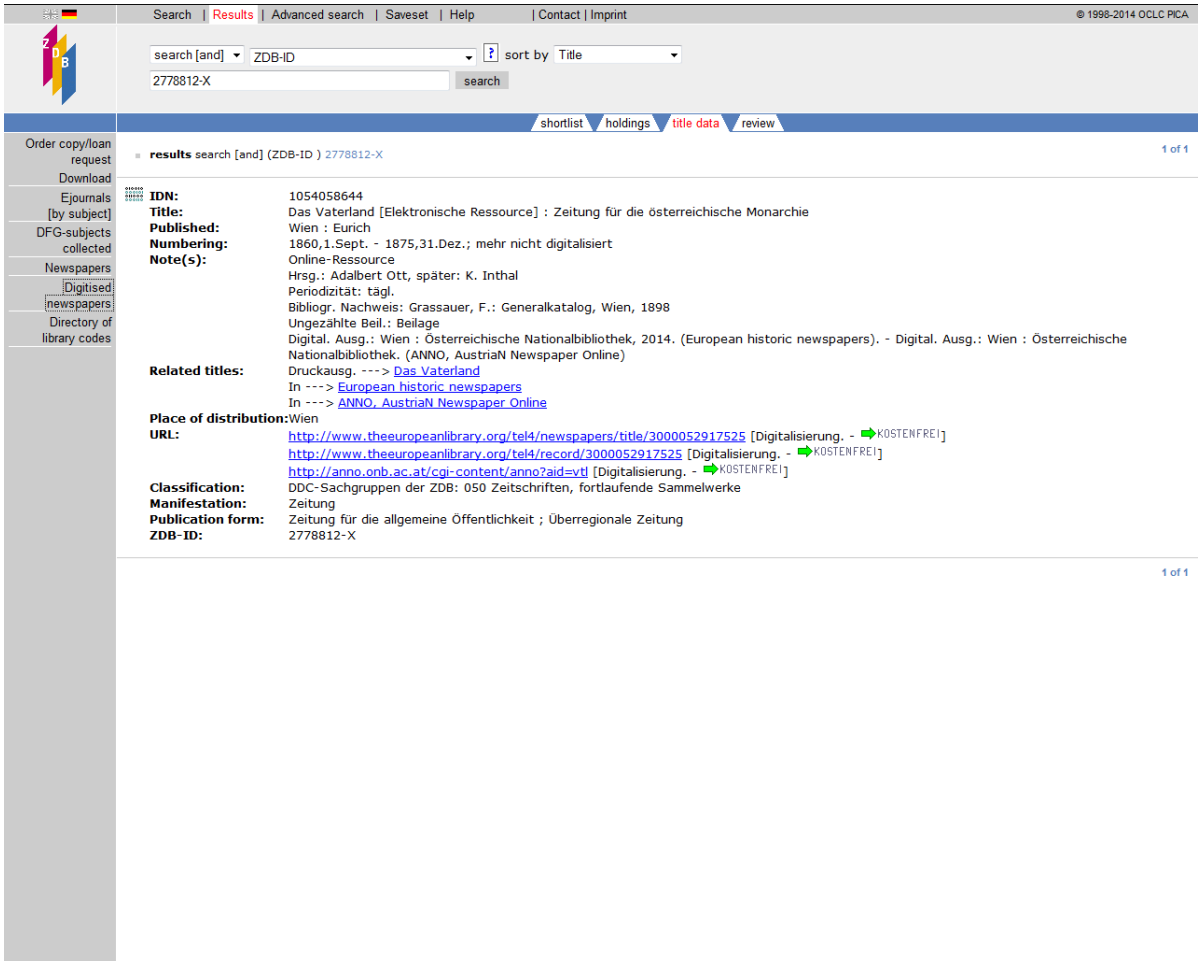

Fig. 3: ZDB-view of a title level record of an Austrian newspaper digitised within the ENP (including links to TEL and the Austrian ANNO-portal)

During the course of the ENP a number of conferences, workshops and information days were organised to enter into discussions with different communities that take an interest in newspaper digitisation. Developed software was presented to and evaluated by technical communities, issues concerning actual workflows were debated with librarians, and the question of access to the digitised newspapers was discussed with prospective users. Talking to members of quite different research communities, one cry could not be overheard: Give us free access to the content! This emphatic appeal was usually followed by the addition: But clean up the texts first, and do it quickly.

In the following I will discuss aspects related to the two questions implicit in the users' requests: What does "free access"[17] actually mean in terms of services provided by libraries, and how do we deal with faulty digital text corpora?

Services

Typically, the newspaper resources refined with OCR, OLR, and NER within the ENP can be reached via four different access points: firstly, there are local presentation environments, usually run by the data providing libraries. Then there is TEL's newspaper

[17] The term "free access" as used here does not have any legal implications. It is assumed that historic newspapers are in the public domain, and that access to their digital representation is free of any restrictions. Copyright issues typically associated with contemporary newspapers are not addressed.

portal, thirdly Europeana, and, last but not least, the ZDB. Europeana and ZDB function as metadata hubs and pointers to the actual digital resources, though Europeana additionally provides access to the actual data via the integrated TEL newspaper browser. Both the local presentation environments and TEL offer all newspaper data of the ENP in its full richness, i.e. including texts, images and mark-ups.

Local newspaper web pages usually offer calendar navigation and, less frequently, options for text searches. Occasionally users can use facets to limit searches and narrow down results, various types of mark-ups and annotations may be available, and links to external information resources may be provided.[18] The most obvious limitation of local presentation environments is their fairly low level of aggregation, usually restricted to newspaper holdings of a specific library.

In many cases the simple principle "the bigger the better" is applicable to the size of digital text corpora. Depending on the type of searches a user wishes to perform, TEL will be a good option since one can search over vast numbers of newspaper issues from all over Europe at the same time.[19] Filtering by title, date, owning library, country, and language is possible. However, it is the mass of available text content that constitutes the most significant feature of the TEL newspaper web page and one that makes it unique for users.

Given the popularity among virtually all user groups and in light of the many possibilities for discovery that large digital text corpora have to offer, it is somehow surprising that libraries, by and large, restrict their online newspaper services to navigation and text searches. There are notable exceptions: The British Library and the KB – National Library of the Netherlands, to name two examples, have created lab environments[20] that invite researchers, developers, and other users to participate in jointly developing tools and services, and working with these on the libraries' digital collections.

It has been stated that digitised historic newspapers attract all sorts of users. It is equally true to say that it is not entirely clear at this point of time what kind of services libraries should provide. The potentials for analysing large digital text corpora are truly exciting. A look at Google's Ngram service[21] provides some insights as to how powerful a plain frequency analysis can be. But libraries can do much better: we have highly reliable information resources with which we can enrich text corpora in a very meaningful way. Our users can greatly benefit from our data being highly structured and therefore easily to be interpreted by machines. The KB lab, though still under development, indicates some possibilities and paths to pursue. One thing is true: Users, librarians, and developers need to gain a much better understanding of the capabilities of natural language processing and text mining techniques, of visualisation methods and intelligent cross-media linking, to name but a few aspects. Useable results will be achieved if libraries tackle these problems together with their users and develop tools and methods that are adaptable to both the material and the actual research questions, in that way creating true benefits. Libraries, being the custodians of

[18] One of the most comprehensive examples of a local presentation environment is ZEFYS and its dedicated pages for three major newspapers of the former German Democratic Republic, see http://zefys.staatsbibliothek-berlin.de/ddr-presse/?no_cache=1. Note that these web pages are available in German only.

[19] Searches in TEL are, in many cases, obviously affected by the multitude of languages of the texts in question.

[20] See http://labs.bl.uk/ and http://lab.kbresearch.nl/

[21] https://books.google.com/ngrams

the digital collections, are the natural place for creating the required organisational and technical infrastructure.

An alternative and often desired form of free access is the provision of the data itself, in a variety of formats, via APIs that support a variety of protocols, and without any restrictions regarding the use of the data. Users will work with the digital resources in their own environments, in universities and research institutions, in commercial businesses and at the home desk, and they will publish and make available their results. It is to be hoped that more and more libraries will make their text data freely available for download and unlimited further use. This data belongs to the public – libraries need to deliver!

Refinement Results

Clearly, the quality of the digital text corpora is fundamental for all types of search and analysis services and largely determines the extent to which these can be applied in a meaningful way. A brief glance at the production conditions is not without sobering effect: In the face of the sheer quantities of historic newspapers kept in libraries, highly automated workflows for the entire digitisation process including OCR, OLR, and other mark-ups have to be put in place, if only for economic reasons. Some specific newspapers of particular interest may well be digitised with great staff effort invested in manual corrections to ensure near perfect results. Such "boutique digitisation", however, needs to be limited to special resources, otherwise great numbers of newspapers will, due to a lack of staff and financial means, remain untouched. When asked whether they prefer "class" or "mass", users will generally opt for the latter approach, too.

Mass digitisation projects will, in many cases, use microfilms as source material. The quality of the microfilms affects the entire workflow significantly. In the best case the grey scale of the microfilm's images adds contrast to a washy print of the original. In the worst case images are skewed or warped, poor binding of the originals results in partial text loss, pages are missing or duplicated or following a wrong order. Elaborate and changing layouts, and the frequent use of Gothic print, "Fraktur", in the originals add complexity to the refinement processes. The manual effort required for the semi-automated post-processing of the digital files, and ultimately the quality of the end results, namely the full texts of the newspaper content, largely depend on these factors.

Digitising newspaper originals involves some of the above plus additional challenges. Newspapers are usually bound in large folio volumes and often the bindings need to be removed to avoid text loss. The actual paper is more often than not brittle and prone to immediate damage.[22] Digitisation projects, even those undertaken in cooperation with a commercial digitiser, can therefore frequently only be executed under difficult conditions on the premises of the library near the actual storage place of the newspapers.

As a result automatically created digital texts are, also depending on the language of the source material, usually of less than satisfactory quality (see Fig. 1). In dealing with this problem libraries need to be creative. Manual corrections performed by library staff or commercial vendors are effective but too expensive to be applied to large collections. Engaging users in the correction process via crowdsourcing can be an option in many cases

[22] Newspapers of the 16th to 18th centuries were mostly published in smaller formats that are easier to handle. The level of paper quality drops with the availability of cheap mass-produced material in the 19th century during which the publication of newspapers became a mass phenomenon.

but it remains dubious whether this approach provides the desired results with respect to less popular resources.

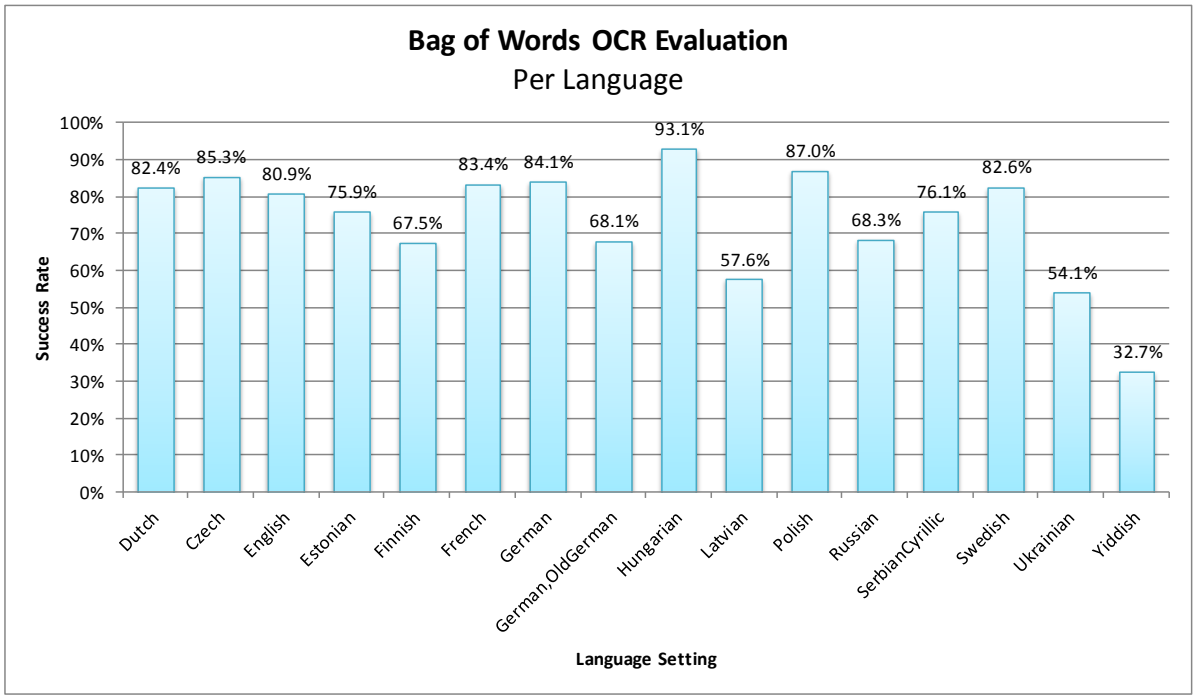

Fig. 4: Bag of words evaluation of OCR results per language[23]

When thinking about corrections it is worthwhile to bear in mind that not all parts of a newspaper are of equal importance. Some text parts that are particularly difficult to process automatically are of limited interest to users, e.g. the rarely changing imprint. It seems advisable, particularly with a view on (semi-)automated corrections supported by software tools, to first concentrate on more significant elements and parts of the texts. Many use cases involve entities such as persons, organisations and places. With the help of NER software such entities may be identified, classified, and linked to other librarian and non-librarian resources, e.g. national authority files, VIAF, Wikipedia, biographical dictionaries etc. In cases where NER produces ambiguous results, software aided support like an auto-suggest function pointing to various authority records could help solving the proverbial "John Smith problem".[24] Taking NER one step further, more abstract entities, e.g. concepts and events, may be included. Disambiguated results can function as "islands of meaning" in the vast text corpora, enabling further services like the visualisation of networks of people, perception of events, ideas, (literary) works etc. Within the ENP NER for Dutch, German and French texts was applied. Obviously there is a need to enable NER for further languages as well as improving its overall functionalities and performance.

[23] The remarkable differences of results per language may either be caused by the complexity of certain languages and scripts or by the trivial fact that languages with fairly small numbers of native speakers simply are not as well supported by OCR technology as more widespread languages that offer a greater commercial potential. Fig. 2 was taken from the "Performance Evaluation Report" of the ENP and describes results of tests run by the project partner University of Salford.

[24] The "Library of Congress Authorities" (http://authorities.loc.gov/) list hundreds of persons of this name.

As was said above, texts created within the ENP are typically available at least at two different access points: local presentation environments and TEL.[25] It is difficult to predict and largely dependent on users' search scenarios which access point they will prefer. Generally speaking, libraries and the European portals should enable corrections in their respective presentation environments. (Semi-)automated corrections of typical OCR reading errors may be cost-effectively executed in the full text indices, thereby simultaneously affecting huge numbers of different newspaper pages. At the same time corrections, particularly those submitted by users, need to be possible on the page level as well.[26]

Correcting texts at different places will obviously create new challenges that need to be addressed. High-performance interfaces between systems are required to handle potentially large data volumes – will the European portals be financially stable enough to provide the required infrastructure? What would be reasonable update intervals? Many versions of a text may be created before the result finally is deemed to be absolutely correct – will libraries need to preserve all or some of these versions and where would these be accessible?[27] Correctors or correcting software may happen to work simultaneously at different places – how can the loss of corrected data be assured?

Answers to these and further questions are not ready to hand. Libraries are well advised to search for them on an equal footing with users, developers and other interested parties to ensure realistic and useable solutions. More than ever we should embrace every opportunity to share and make freely available our data – and learn from our users.

[25] The data in question may, in fact, be available via many more web pages, e.g. at national aggregation portals like the "Deutsche Digitale Bibliothek"(German Digital Library, see: https://www.deutsche-digitale-bibliothek.de/) in Germany.

[26] The TROVE Digitised Newspapers' correction tool may serve as a good example for an environment that is easy to handle for users. For the impressive results see https://trove.nla.gov.au/ndp/del/hallOfFame. Comparable tools are available via the KB lab.

[27] Changing texts create a profound problem in the context of scientific research, where the traceability of citations and references is of indispensable value.