Digitizing University Libraries – Evolving from Full-Text Providers to CLARIN Contact Points on Campuses

Manfred Nölte
State and University Library
Bremen, Germany
noelte@suub.uni-bremen.de

Martin Mehlberg
State and University Library
Bremen, Germany
martin.mehlberg@suub.uni-bremen.de

Abstract

Based on the example of the State and University Library Bremen (SuUB) we will outline in this paper, how academic libraries with digitization activities (hereinafter referred to as *digitizing libraries*) could establish even closer ties to CLARIN in the future. After describing SuUB's past and current CLARIN-related activities (especially full-text transfers to a CLARIN-D centre) we suggest that this collaboration could be expanded by providing advice and training for researchers of the Digital Humanities as potential CLARIN users. Equally important from our point of view is the discussion about future structural options on the level of research infrastructures. We suggest a collaboration between digitizing libraries to jointly agree upon standards of data quality, file formats, interfaces and web services. We discuss the foundation of local CLARIN contact points to pass scholars and researchers on to the respective contact or service of CLARIN. The relevance to CLARIN activities, resources, tools or services is described at the end of each respective section. From the conclusions, the reader will notice: It is the right time for change.

1 Digitizing University Libraries as Full-Text Providers for CLARIN

The State and University Library Bremen is one of many libraries dedicated to the digitization of its historical collections. Digitization and especially the generation of full text is an important instrument for improving the accessibility of valuable information contained in fragile historical documents. It facilitates academic research and teaching and is indispensable to the Digital Humanities. By doing so, these libraries play a very important role as full-text providers or creators of data.

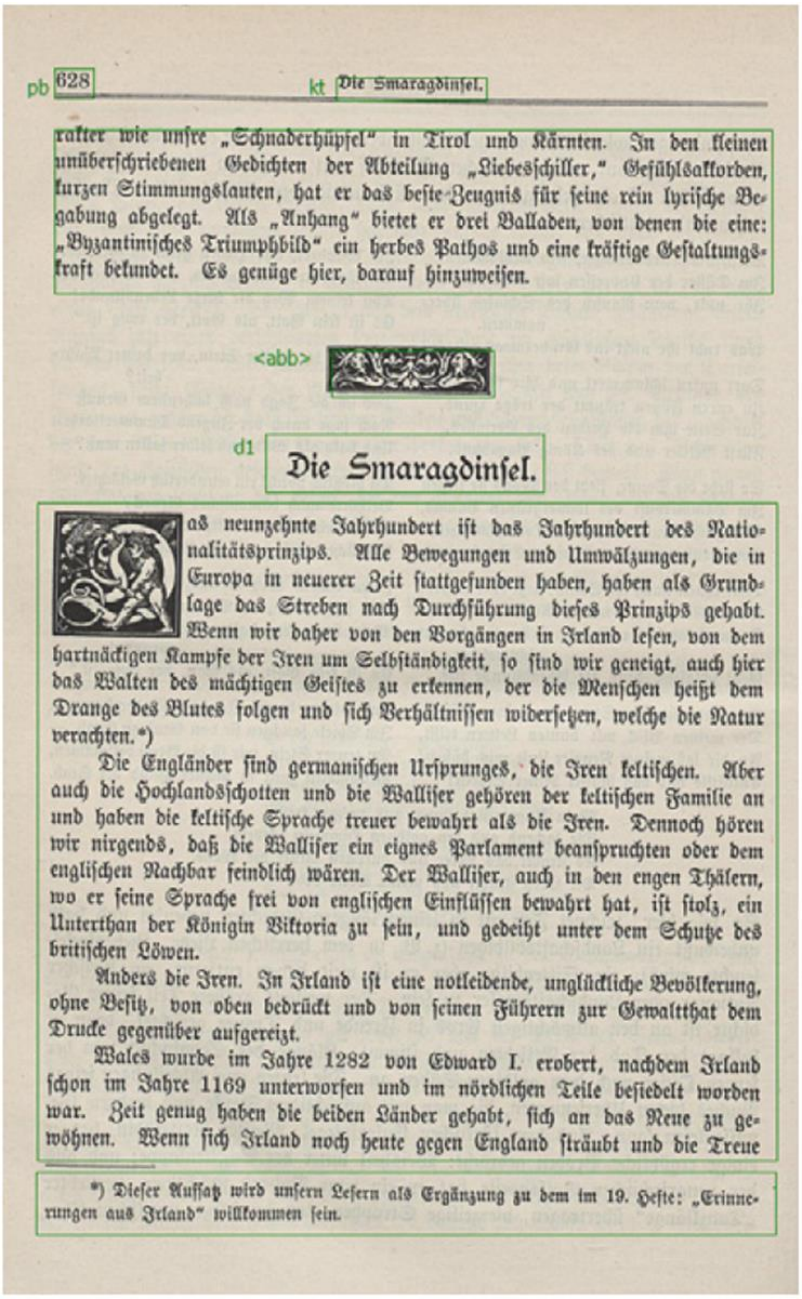
Usually, university libraries undertaking digitization projects produce digital images, metadata for cataloguing and web-navigation purposes, and optical character recognition (OCR) full text for searching. These resources are made available through the library's web portal for digital collections. However, digital humanists need rather high-quality full texts enriched with metadata in the appropriate format in order to process and analyze them with powerful software tools (like regular expression search, part-of-speech tagging, named-entity recognition or topic modeling). To satisfy this specific demand, the SuUB has actively transferred full texts created through digitization projects (funded by the German Research Foundation, DFG) to the Berlin CLARIN-D centre and thus via metadata harvesting to the CLARIN research infrastructure. All these CLARIN tools, concentrated, documented and ready to use, have been a great motivation for this activity. We would like to outline our approach adopted so far, the results and the dissemination achieved within the scientific community. Later on, in section 3 we discuss the underlying structure and concept and how we might intensify operations like this.

The historical journal *Die Grenzboten* was the first full text transferred from the SuUB to CLARIN (Geyken et al., 2018). *Die Grenzboten* is a long running serial publication (1841–1922) which can be classified as a literary journal that also covers politics and arts. It was founded by Ignaz Kuranda (1811-1884) in Brussels in 1841 and later on published in Leipzig and Berlin. We demonstrate that good OCR quality and a page-by-page structuring are prerequisites for the creation of a high-quality

Manfrend Nölte and Martin Mehlberg 2021. Digitizing University Libraries - Evolving from Full-Text Providers to CLARIN Contact Points on Campuses. *Selected papers from the CLARIN Annual Conference 2020*. Linköping Electronic Conference Proceedings 180: 180 155–164.

Text Encoding Initiative (TEI) version of a full text. The TEI version was created in cooperation with the Deutsches Textarchiv (DTA) at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW) (Nölte et al., 2016).

We digitized more than 185,000 single pages in 270 volumes. Almost 33,000 articles were digitized via optical character recognition (OCR) and the titles of the articles were manually captured. The resulting OCR full text was processed by the OCR software ABBYY Finereader 9 and consists of approximately 500 million characters and 65 million tokens. As a second aspect of text quality, we enhanced the level of document structure according to an agreed standard format together with our partners, the Deutsches Textarchiv (DTA; Haaf Geyken and Wiegand, 2014/15). Figure 1 shows manually corrected and tagged "zoning information" based on coordinates provided by the ABBYY Finereader XML files. Using this structure information, we converted the OCR output format to an interoperable TEI format. The metadata of the 33,000 articles also contain information about the publication dates, so that it is possible to analyze the full texts over time.


pb 628 kt Die Smaragdinsel.

rakter wie unsre „Schnaderhüpfel" in Tirol und Kärnten. In den kleinen unüberschriebenen Gedichten der Abteilung „Liebesschiller," Gefühlsakkorden, kurzen Stimmungslauten, hat er das beste Zeugnis für seine rein lyrische Begabung abgelegt. Als „Anhang" bietet er drei Balladen, von denen die eine: „Byzantinisches Triumphbild" ein herbes Pathos und eine kräftige Gestaltungskraft bekundet. Es genüge hier, darauf hinzuweisen.

<abb>

d1

Die Smaragdinsel.

Das neunzehnte Jahrhundert ist das Jahrhundert des Nationalitätsprinzips. Alle Bewegungen und Umwälzungen, die in Europa in neuerer Zeit stattgefunden haben, haben als Grundlage das Streben nach Durchführung dieses Prinzips gehabt. Wenn wir daher von den Vorgängen in Irland lesen, von dem hartnäckigen Kampfe der Iren um Selbständigkeit, so sind wir geneigt, auch hier das Walten des mächtigen Geistes zu erkennen, der die Menschen heißt dem Drange des Blutes folgen und sich Verhältnissen widersetzen, welche die Natur verachten.*)

Die Engländer sind germanischen Ursprunges, die Iren keltischen. Aber auch die Hochlandsschotten und die Walliser gehören der keltischen Familie an und haben die keltische Sprache treuer bewahrt als die Iren. Dennoch hören wir nirgends, daß die Walliser ein eignes Parlament beanspruchten oder dem englischen Nachbar feindlich wären. Der Walliser, auch in den engen Thälern, wo er seine Sprache frei von englischen Einflüssen bewahrt hat, ist stolz, ein Unterthan der Königin Viktoria zu sein, und gedeiht unter dem Schutze des britischen Löwen.

Anders die Iren. In Irland ist eine notleidende, unglückliche Bevölkerung, ohne Besitz, von oben bedrückt und von seinen Führern zur Gewaltthat dem Drucke gegenüber aufgereizt.

Wales wurde im Jahre 1282 von Edward I. erobert, nachdem Irland schon im Jahre 1169 unterworfen und im nördlichen Teile besiedelt worden war. Zeit genug haben die beiden Länder gehabt, sich an das Neue zu gewöhnen. Wenn sich Irland noch heute gegen England sträubt und die Treue

*) Dieser Aufsatz wird unsern Lesern als Ergänzung zu dem im 19. Hefte: „Erinnerungen aus Irland" willkommen sein.

Figure 1: "Zoning" – adding structural annotation to the pages

We were active to disseminate our digitized historical journal enhancing its outreach. It has been used as textual material by computer linguists, digital humanists and philologists, as well as being part of

academic teaching at universities like Ghent, Würzburg and Göttingen[1]. Together with ground truth full text data, it has been used by OCR post correction system providers like PoCoTo (Vobl et al., 2014) and ProjectComputing (Evershed and Fitch, 2014) and big projects like OCR-D (Neudecker et al., 2019). The journal *Die Grenzboten* was subject to diachronic collocation analyses[2] (Jurish and Nieländer, 2020) as well as to analyses like topic modeling (Jannidis, 2016; Fechner and Weiß, 2017; Graham et al., 2012). The project website at SuUB lists five other research projects in the context of the humanities.

It is clear that this usage is a result of actively providing the journal full text to the scholarly community and to research infrastructures as described in this paper. Up to now this journal is just an example of a serial source available at this level of quality and accessibility. A lot more have been digitized or are in the process of being digitized. Later on for example, the SuUB and the University Library Frankfurt digitized over 1000 book titles with about 245,000 pages within the project "Digitale Sammlung Deutscher Kolonialismus" (a digital collection of texts from the period of German colonialism). This project also generated full texts that have been transferred to CLARIN. Both full-text transfers are showcases for a collaboration/teamwork between a digitizing library and CLARIN. The question now is: What has to be done to intensify the transfer of huge amounts of digitized full texts in a reasonable and cost-effective routine manner? We will revisit this question in more detail in section 3.

In the following part we describe the relevance of the issues mentioned above for the CLARIN activities. As described above, the SuUB has actively transferred full texts to a CLARIN-D centre in an early phase directly after the digitization process. Doing so, we helped to increase the amount of language resources provided by CLARIN and together we generated a bigger perception of this full text within the scientific community. CLARINs "ingest service" and the possibility to host the full text in the CLARIN repositories also clearly helped to increase the dissemination of our full texts within the scientific community.

Digitization activities have the potential to create huge amounts of digitized full texts. This in turn stimulates inter- and cross-disciplinary research. This first collaboration between CLARIN and the SuUB Bremen can be seen as a showcase scenario of how content-providing libraries and CLARIN can mutually benefit from these kinds of digitization projects.

Like CLARIN, we seek to take the requirements of the user community into consideration. For a library it is business as usual to have a lot of contact with our "users" (patrons), especially if they are interested in the [digitized] material we hold. This way we know the demands and expertise of our library patrons.

With our digitized full-text resources we have been serving a lot of different user communities: German philology, linguists, Digital Humanities, political science, history, etc.

2 Counselling and Training Activities of the SuUB Bremen

Academic libraries are close to scholars and researchers not only in terms of physical closeness but also in terms of subject proximity (providing information and services). The main task of libraries has been (and still is) the supply of literature/scientific information. All libraries have contact points and recently they have even gained importance on many campuses, because they provide a "learning space" for students and scientists.

Libraries of course know "their own material", i.e. texts digitized by themselves, best. They know the subject, the context and the quality of the digitized material. The latter manifests itself in the

[1] The journal *Die Grenzboten* as research data on GitHub; 'Expertenworkshop: Topic Modelling' at the University of Göttingen (May 2018); a workshop by Bryan Jurish, Thomas Wernecke, and Maret Nieländer on 'Diacollo and *Die Grenzboten*: Exploring Diachronic Collocations in a Historical German Newspaper Corpus' at the Genealogies of Knowledge I — Translating Political and Scientific Thought across Time and Space conference at the University of Manchester (December 2017); 'CIS OCR Workshop v1.0: OCR and Post Correction of Early Printings for Digital Humanities' at the Ludwig Maximilian University of Munich (LMU) (September 2015); and a module, also at LMU, by Florian Fink on *PoCoTo: Practice* (2015) [all accessed 8 October 2018].

[2] Jurish, Bryan, M. Nieländer, and T. Werneke. 2017. "DiaCollo and *die Grenzboten*." Talk presented at the conference GenealogJurish, B., M. Nieländer, and T. Werneke. "DiaCollo and die Grenzboten." Talk presented at the conference Genealogies of Knowledge I: Translating Political and Scientific Thought across Time and Space, University of Manchester, 7th-9th December, 2017ies of Knowledge I: Translating Political and Scientific Thought across Time and Space, University of Manchester, 7th-9th December.

condition of the original material, pixel images, error rates of the full texts as well as the scope and the quality of the metadata. Technical issues, for example, are the interfaces to get access to the data and the formats delivered by the respective systems. In summary, this enables libraries to help with quality issues, access related issues, even content-related issues or options to use web services like IIIF[3] (Snydman et al., 2015). For example, the quality of the section titles within the full text of *Die Grenzboten* is poor, due to the usage of a special character type, whereas the same information contained in the METS-XML files has perfect quality, because it was captured manually.

In the past, the SuUB has gathered some experience with the personal counselling of scholars from across the humanities disciplines, including linguistics and political science. In general, the questions were of a technical nature (relating for example to formats or system interfaces), but sometimes also questions of a more theoretical nature were discussed (e.g., kinds of quantitative analyses like topic modeling, diachronic collocations, etc.).

The aim of the technical advice was primarily to provide researchers with the knowledge necessary to make full texts available in an interoperable format that meets the requirements of specific software tools. Especially with structured full texts (TEI or in general all sorts of XML), format issues have to be considered. Some quantitative tools, like mallet (topic modelling; Graham et al., 2012) only need plain text. But the pre-processing or the whole tool chain (e.g., including a graphical presentation for the analysis findings) nearly always requires the above-mentioned features: structured pages (i.e., semantically tagged full text) and metadata (year of publication, authors, etc.).

Another point of view is to consider the several target groups (like bachelor, master and Ph.D. students, postdocs, researchers, lecturers and citizen scientists) within the library patrons appropriately. What target group is supposed to be passed on to CLARIN? What level of counselling are libraries able to fulfil? We need to find and define a certain transfer point.

In the following part we describe the relevance of the issues mentioned above for the CLARIN activities. As shown, digitizing libraries are in a good position to start researcher training activities with respect to their full-text resources. Furthermore, they can help access web services or metadata offered by the digital collections software systems, like IIIF and OAI-PMH.

Actively supporting the above-mentioned full-text transfers and the mentioned counselling activities will result in considerably better outcomes in all fields of automated and computer-aided research across disciplines working with digitized material. It will enable the employment of quantitative methods and approaches such as authorship attribution studies, clustering techniques (i.e., for literary genre analysis), topic modeling etc.

3 Prospects for Future Collaboration Between CLARIN and Academic Libraries

As shown above digitizing libraries already play a role in the context of CLARIN and the group of CLARIN users. The next step should be to intensify the collaboration between CLARIN and those libraries in order to harmonize the provisioning or transfer of digital textual material, to jointly agree upon common activities or to even establish CLARIN contact points on university campuses. The most appropriate place for these contact points is a library as we will demonstrate in the next section.

[3] An IIIF-manifest enables an IIIF-viewer to establish a direct metadata link to the respective digitized resource. Here a viewer on universalviewer.io links to the SuUB Bremen:
`http://universalviewer.io/uv.html?manifest=https://brema.suub.uni-bremen.de/i3f/v20/1702471/manifest#?c=0&m=0&s=0&cv=6&xywh=-981%2C-124%2C4978%2C2444`

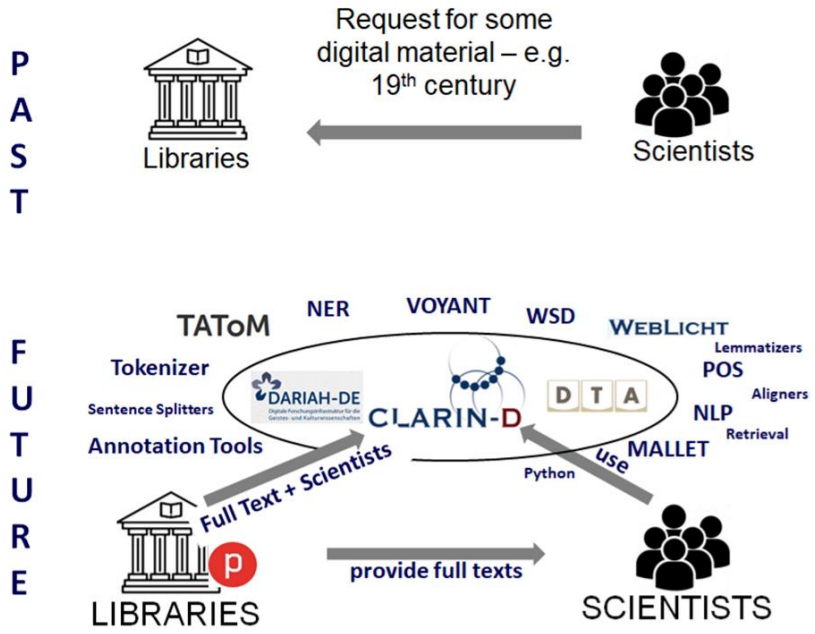


Figure 2: Prospects for Future Collaboration between CLARIN and Academic Libraries

Having done full-text transfers, mentioned in section 1, a few times we list some criteria that might be in need of a more precise specification together with potential requirements. Here we give only short explanations, see (Nölte and Blenkle, 2019) for more examples and details.

- *Full-text quality*: For example, a maximum error rate of characters or for other elements of the metadata, like the structuring of the text. These criteria may vary for different centuries or decades or different software tools or scientific approaches.
- *File formats and metadata*: Transferring plain text is not an option, nor is the output of OCR engines (such as ABBYY-XML). There has to be a decision for ALTO, PAGE, TEI or other file formats, possibly together with 'annotation guidelines'. (Haaf et al., 2014/15)
  Supplemental note: The use of format converters should be considered with care. There will always be a loss of information converting from a format to another. A good and simple test might be to convert from format *A* to format *B* and back to *A*, and to compare the emerging differences.
- *Persistent back links*: Within the full text there should be back links (page-by-page or at least by sections) to the scanned images in the digital collections of the respective library or archive. If possible, these back links should be persistent at a page "URN granular" level (Sommer, 2010)[4]. Researchers appreciate having the possibility to check the original image quality or to have access to supplemental material such as graphics, images, advertisements, or vignettes.
- *Line breaks*: There should be a guideline whether to transcribe line breaks as is. We have cooperated with partners with varying opinions on this question. Some institutions wanted line breaks as is, whilst the transcription for Wikisource had to be without wrapped words.[5]
- *Strictness of character transcription:* Within historical full texts the spelling, of course, should be transcribed as is; for instance, 'Säugethiere' with 'th' and 'Entwickelung' instead of the modern form 'Entwicklung'. The same should apply to the transcription of single historical characters using UTF8 codes, like ligatures or special historical glyphs. However, some tools (especially quantitative tools) or workflows may well require transliterated standard versions of the special characters.

[4] There has been a pilot project for the persistent identification for individual pages, but up to now the registration of massive numbers of URNs is not common practice within digitization activities at libraries.
[5] `https://de.wikisource.org/wiki/Die_Grenzboten`

Figure 3: Two examples of historical characters; left: a historical version of an umlaut-a; right: the long-s

Reflecting on these criteria also helps to remind the researcher of the original source of the material. With an OCRed full text being the 'model' of a paper original, Piotrowski (2019) mentioned a 'mapping property' and a 'reduction property' of models that researches should be aware of. For example, the analyses of serial sources should also consider basic textual properties. An example is the distribution of text quality over time (i.e. year of publication), which might have an impact on methods used in the Digital Humanities. It is therefore necessary to develop methods with a stability[6] towards varying OCR error rates. Methods and models requiring a constant error rate should be adapted or it should be possible to critically assess and interpret the results. Similarly, some of the mentioned criteria (such as 'line breaks' and 'strictness of character transcription') will also have an impact on the need for adaptation or interpretation of the used methods or an adapted pre-processing of the full texts.

Ideally, there should be documentation listing all the above-mentioned information: the level of the 'full-text and metadata' quality, whether there are further file formats available, the availability of back links, and the status of line breaks and character transcriptions. If this 'full-text metadata' is realized with computer readable XML formats, pre-processing scripts might automatically decide what pre-processing remains to be done, and what analysis tools or scientific approaches might be applicable. Licensing and intellectual property would be a further major issue to address.

Jointly discussing and agreeing upon criteria and requirements like this will lead to a best practice approach for future transfers of full texts to CLARIN. Together we might even go for a fully automated full-text transfer or harvesting as a long-term goal. And the above mentioned 'full-text metadata' might partially automate the preselection of texts to process.

In the following part we describe the relevance of the issues mentioned above for the CLARIN activities. Here we have also addressed the issue of data quality. OCRed full-text resources have a certain error rate, and metadata may be rich and good or sparse and of poor quality. Setting up the described collaboration will standardize and harmonize all future full-text transfers and training activities in the context of CLARIN and digitizing libraries, i.e., together we will create best practice approaches that provide scholars and researchers with the best possible quality and interoperability of language resources and services.

4 Prospects for Future CLARIN Contact Points on Campuses

As proposed in section 3, another future activity is the establishment of CLARIN contact points for scholars and scientists at academic libraries. These already have a proficiency in counselling and offering services in the respective domains. They also function as learning spaces, aiming at creating the best atmosphere for the exchange of knowledge. Other outstanding advantages of libraries are: Libraries constitute sustainable structures in the scientific world, they are research infrastructures themselves and they are local on the campus. Every university has a library and universities are places where students become scientists who might be passed on to CLARIN as potential users.

Currently there are a lot of activities to establish event formats at libraries as scholar or researcher training activities with names like "digital lab", "hands-on lab", "GLAM[7] lab", "innovation lab", "data lab"[8], "HackyHours"[9], "Digital Learning Lab"[10], "Digital Humanities lab"[11], "Library Labs"[12],

[6] The 'stability' of an algorithm or method refers to the quality of the results of a stable method that does not degrade (that much) whilst the given input has a reduced or degraded quality.
[7] GLAM is an acronym for "galleries, libraries, archives, and museums".
[8] DataLab, https://www.uni-goettingen.de/de/daten+lesen+lernen/592287.html
[9] HackyHours, https://librarycarpentry.org/blog/2019/06/hackyhours-zbmed/
[10] Digital Learning Lab , https://www.uni-marburg.de/de/ub/lernen/kurse-beratung/wissen-organisieren/dll
[11] Digital Humanities Lab, https://ub.fau.de/forschen/digital-humanities-lab/
[12] Library Labs, https://www.bl.uk/projects/british-library-labs#

"scholarly makerspaces", "Digital Literacy"[13] and more combinations of these words. CLARIN might play a role to harmonize this multitude of activities, to combine these with CLARIN's experience, services and tools. The above-mentioned start of collaboration between libraries and CLARIN might be a good first step.

To establish CLARIN contact points on Campuses in the most appropriate manner, the needs/demands/requirements of the scholars and the current situation of academic libraries have to be taken into account. What is on the scholars' wish lists? What concrete services would researchers like to see? What are libraries able to fulfil? The VDB[14] annual report of the commission for research-oriented services (Leiß, 2020) states: Scientists want sample solutions and best-practice solutions for typical use cases in order to identify and omit problems early. And there is already a professional qualification and knowledge of typical difficulties and pitfalls together with their respective solutions. Scientists want central services and contact persons, which libraries offer.

But of course, something is new. The digital change has come into play. As we described above, services, book titles and pages have already partially turned into web services, metadata and files of different formats. This, for sure, will have an impact on new job profiles within academic libraries. A new strategy for recruiting and professional training will be necessary. Libraries have historically found and will in the future find a level to cope with the multitude of requirements, topics and user communities to accomplish a good start regarding the proposed activities. Another helpful approach is the collaboration between libraries. We have participated within YUFE[15] network activities for an exchange of experience. From the library of Maastricht we have learned that a specific service needs time for being accepted and for being recognized and well used. This means that libraries need to increase outreach efforts to maximize the dissemination of their services. Last but not least, an integration of these new library services into the curricula of the respective departments of the university will help to stabilize these services and the "flow" of participants right from the lectures to the libraries and finally to CLARIN.

In the following part we describe the relevance of the issues mentioned above for the CLARIN activities. Relevance to the CLARIN activities: As shown, the SuUB together with CLARIN has a big potential to establish user assistance, a help desk or contact point. While documenting the digital collections software system with user manuals, we might support scholars and researchers within the domain of digital language resources. An example is information about our systems' persistent identifiers and citation mechanisms (see the above criterion for "Persistent back links").

Contact points on Campuses also might be a useful activity of an academic library to increase the awareness of useful tools and resources of CLARIN.

5 Conclusions

Here we describe conclusions for scholars of the Digital Humanities and for data providing institutions, i.e., for CLARIN, libraries and the whole GLAM sector. As the authors come from the professional context of a digitizing library, they refer to this type of institution in the list of conclusions. Some or all of the conclusions may also be applicable to the entire GLAM sector or to all data-providing institutions, as indicated or in a modified form. The following list is a summary of statements discussed in the previous sections. Furthermore, invitations are given to start or intensify joint activities (in italics).

5.1 Conclusions for CLARIN and digitizing libraries

- *Get together and coordinate.*
  Together we should setup a network to streamline the collection and propagation of the requirements of the user community back to the data creation institutions and to address the issues of data quality and data completeness.
  We should standardize and harmonize all future full-text transfers and training activities in the context of CLARIN and digitizing libraries. Libraries are in a good position to start researcher

[13] Digital Literacy, https://www.tu-braunschweig.de/lehre/konzepte-tools-und-projekte/future-skills#c647807

[14] VDB: Association of German Librarians (Verein Deutscher Bibliothekarinnen und Bibliothekare, https://www.vdb-online.org/)

[15] YUFE – Young Universities for the Future of Europe, https://yufe.eu/

training activities with respect to their full-text resources. They are used to address diverse scientific communities; they are local on the campus and are mostly already well equipped with learning spaces. This way we will create best practice approaches that provide scholars and researchers with the best possible quality and interoperability of language resources and services.

- *Help with establishing CLARIN contact points on campuses*
  Scholars and researchers might be passed on to CLARIN or directly to the respective contact or service of CLARIN. There are a few questions to consider. What target group is supposed to be directed to CLARIN (Potential target groups: students (bachelor, master, Ph.D.), postdocs, researchers, lecturers and citizen scientists)? What level of counselling are libraries able to fulfil? We need to find and define a certain transfer point.
- *Consider each other as partners*
  Libraries help to increase the amount of language resources provided by CLARIN and might also help to increase the awareness of useful tools and resources of CLARIN. CLARIN helps with the possibility to host the full text in the repositories with the "ingest service". It also increases the dissemination of full texts within the scientific community.
- *Integrate the whole GLAM sector*
  Together with data providing institutions inter- and cross-disciplinary research will be stimulated in the best possible manner.

Finally, there are some specific conclusions for libraries.

- *Continue to develop*
  As we have learned from contacts to the above mentioned YUFE network, there are already libraries with new positions and job titles like data steward, data specialist, information specialist [digital] humanities, research data manager or education and research technician. Libraries should establish further professional training and should create those new positions, to meet new requirements. A further strategy might be to adapt the recruiting policy.
- *Pay attention to CLARIN activities, services and tools*
  It should be clearly seen what CLARIN has achieved so far and how fast services and tools around digital resources have changed. Libraries should keep up or even better engage with existing and upcoming activities in the domain of digital research infrastructures.
- *Collaborate with other libraries or institutions from the GLAM sector*
  The best way for an exchange of experience is by collaboration. With regard to approaches of digitization of research and education, locality and regionality are playing an increasingly minor role anyway.
- *Streamline IPR issues*
  Intellectual property rights (IPR) should be kept as simple, open and transparent as possible.

5.2 Conclusions for scholars of the Digital Humanities

- *Help and engage yourself*
  Help us to actively shape and build the above-mentioned training activities, and in the long run CLARIN contact points.
- *Consider academic libraries as helpful places*
  Academic libraries will continue to provide huge amounts of digital content, respectively digital full texts. Libraries have valuable corpora for historical research. We suggest that researchers should regard libraries (and research infrastructures) even more intensively as partners to get access to historical full-text materials. Everybody might agree that digitizing libraries contribute significantly to the amount of the available digitized material. Still, a novel approach is to actively transfer (or provide via metadata harvesting) full texts to the researchers and to research infrastructures, which is one of the central issues of this paper.
  Libraries have always been helpful with knowledge of "their" library stock (collections, literary remains, manuscripts, etc.). More and more they also help to access web services or metadata offered by the digital collections software systems, like IIIF (Snydman et al., 2015) and OAI-PMH.
- *Use "digitization on demand" services*
  Help with increasing the amount of digitized full-text resources.

- *Know about the resources you use*
  E.g., be aware of the original material and the whole process it has undergone (see the list of criteria in section 3).
- *Please give feedback*[16]
  Tell us about your work with regards to the material you have been using. Contact the respective library or CLARIN (these both should intercommunicate anyway). Tell whether the material has matched your requirements with respect to the criteria discussed in section 3.

It is the right time for change, for new collaborations, new structures and powerful digital research infrastructures.

Acknowledgements

We would like to thank the anonymous reviewers for the helpful comments and suggestions. This work has benefited a lot from the results of several earlier projects funded by the German Research Foundation (DFG).[17]

References

Evershed, John and Kent Fitch. 2014. 'Correcting Noisy OCR: Context Beats Confusion', in: *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 45–51 (New York: ACM). http://dx.doi.org/10.1145/2595188.2595200.

Fechner, Martin and Andreas Weiß. 2017. 'Einsatz von Topic Modeling in den Geschichtswissenschaften: Wissensbestände des 19. Jahrhunderts', in: *Zeitschrift für digitale Geisteswissenschaften*. DOI: 10.17175/2017_005.

Geyken, Alexander, Matthias Boenig, Susanne Haaf, Bryan Jurish, Christian Thomas, and Frank Wiegand. 2018. Das Deutsche Textarchiv als Forschungsplattform für historische Daten in CLARIN. In: Henning Lobin, Roman Schneider, Andreas Witt (Hgg.): Digitale Infrastrukturen für die germanistische Forschung (= Germanistische Sprachwissenschaft um 2020, Bd. 6). Berlin/Boston, 219–248. Online-Version, DOI: 10.1515/9783110538663-011.

Graham, Shawn, Scott Weingart, and Ian Milligan. 2012. 'Getting Started with Topic Modeling and MALLET'. DOI: 10.46430/phen0017

Haaf, Susanne, Alexander Geyken, and Frank Wiegand. 2014/15. 'The DTA "Base Format": A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources', *Journal of the Text Encoding Initiative*, no. 8, no page.

Jannidis, Fotis. 2016. 'Quantitative Analyse literatischer Texte am Besipied des Topic Modeling', *Der Deutschunterricht*, 68.5, 24–35.

Jurish, Bryan and Maret Nieländer. 2020. Using DiaCollo for Historical Research. Selected papers from the CLARIN Annual Conference 2019. *Linköping Electronic Conference Proceedings* 172: 172 33–40.

Leiß, Caroline. 2020. Bericht der Kommission für forschungsnahe Dienste 2019. O-Bib. *Das Offene Bibliotheksjournal / Herausgeber VDB*, 7(4), 1-3, DOI: 10.5282/o-bib/5628 .

Neudecker, Clemens, Konstantin Baierer, Maria Federbusch, Kay-Michael Würzner; Matthias Boenig, Elisa Herrmann, and Volker Hartmann. 2019. OCR-D: An end-to-end open-source OCR framework for historical documents, in: *Proceedings of the 3rd International Conference on Digital Access to Textual*

[16] The following question has been a discussed within the session "Data Curation, Archives and Libraries" at the virtual CLARIN 2020 conference: *"How to get feedback on whether the work is **deployed** by the community and **if it actually is meaningful**?"* First, **every imaginable type** of feedback would definitely be appreciated by all data providing institutions. A **late feedback** would be having a look at publications on the respective work. We had such positive feedback. In the past, we had direct contact to the users, but on a large scale that is definitely not possible. With respect to the question **"if it actually is meaningful"**: We suggest the collection and forwarding of whether the material was matching the requirements or the type of **further** requirements, to establish some kind of backpropagation through this network of research infrastructures.

[17] DFG funded projects: https://gepris.dfg.de/gepris/projekt/196492153?language=en (two projects), https://gepris.dfg.de/gepris/projekt/324473798

Cultural Heritage, Brüssel 09.05.2019, 53–58. https://dl.acm.org/doi/10.1145/3322905.3322917 [accessed 27 April 2020].

Nölte, Manfred and Martin Blenkle. 2019. '*Die Grenzboten* on its Way to Virtual Research Environments and Infrastructures', *Journal of European Periodical Studies*, 4.1, 19-35.

Nölte, Manfred, Jan-Paul Bultmann, Maik Schünemann, and Martin Blenkle. 2016. 'Automatische Qualitäts-verbesserung von Fraktur-Volltexten aus der Retrodigitalisierung am Beispiel der Zeitschrift *Die Grenzboten*', o-bib, 3.1, 32–55 (p. 32) [accessed 27 April 2020].

Piotrowski, Michael. 2019. 'Historical Models and Serial Sources', *Journal of European Periodical Studies*, 4.1

Snydman, Stuart, Robert Sanderson, and Tom Cramer. 2015. 'The international image interoperability framework (IIIF): a community & technology approach for web-based images', in: *Archiving Conference*, vol. 2015, 16–21. Society for Imaging Science and Technology.

Sommer, Dorothea. 2010. 'Persistent Identifiers: the 'URN Granular' Project of the German National Library and the University and State Library Halle', LIBER Quarterly, 19.3-4, 259–274. DOI: http://doi.org/10.18352/lq.7965 [accessed 26 January 2021].

Vobl, Thorsten, Annettee Gotscharek, Uli Reffle, Christoph Ringlstetter, and Klaus U. Schulz. 2014. PoCoTo - an Open Source System for Efficient Interactive Postcorrection of OCRed Historical Texts. In Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, 57–61. DATeCH '14. New York, NY, USA: ACM. DOI:10.1145/2595188.2595197.