Article

DivaDesk: A Holistic Digital Workspace for Analyzing Historical Document Images

Nicole Eichenberger, Angelika Garz, Kai Chen, Hao Wei, Rolf Ingold, and Marcus Liwicki | Fribourg

Abstract

In this article we present the concept of DivaDesk – a Virtual Research Environment (VRE) for scholarly work on historical documents inspired by the shift toward working with digital facsimiles. The contribution of this article is three-fold. First, a review of existing tools and projects shows that a holistic workspace integrating the latest outcomes of computational Document Image Analysis (DIA) research is still a desideratum that can only be achieved by intensive interdisciplinary collaboration. Second, the underlying modular architecture of the digital workspace is presented. It consists of a set of services that can be combined according to individual scholars' requirements. Furthermore, interoperability with existing frameworks and services allows the research data to be shared with other VREs. The proposed DivaDesk addresses specific research with historical documents, as this is one of the hardest cases in computational DIA. The outcomes of this paradigmatic research can be transferred to other use cases in the humanities. The third contribution of this article is a description of already existing services and user interfaces to be integrated in DivaDesk. They are part of ongoing research at the DIVA[1] research group at the University of Fribourg, Switzerland. The labeling tool DivaDia, for example, provides methods for layout analysis, script analysis, and text recognition of historical documents. These methods build on the concept of incremental learning and provide users with semi-automatic labeling of document parts, such as text, images, and initials. The conception and realization of DivaDesk promises research outcomes both in computer science and in the humanities. Therefore, an interdisciplinary approach and intensive collaboration between scholars in the two research fields are of crucial importance.

1. Introduction

The technological developments seen in the last few decades have triggered a shift in how scholars in the humanities work when viewing historical documents in their repositories whenever research questions require an analysis of the original documents. Thanks to digitization, they are now able to work with digital facsimiles, so access to visual representations of historical documents has become much easier. The increasing amount of digital data available in virtual libraries, such as *e-codices*[2] and *manuscripta mediaevalia*,[3] provides new research possibilities, such as comparing digital facsimiles of different repositories and annotating digital images. In order to handle the data and to perform research tasks on digital facsimiles, scholars need usable tools. In addition to viewing facsimiles, direct searches for specific text passages in the digitized data are also needed. Furthermore, linking research data and annotations with corresponding text passages would be highly beneficial. Another desideratum is a tool for viewing all the samples of a specific text phrase, initial, or decoration contained in a document (or set of them).

While tools for specific tasks[4] do exist, none of them serve scholars in the humanities in all aspects of their work on digital facsimiles, i.e. the generation and presentation of item descriptions, content representation, and research data. Our vision is to realize DivaDesk, a VRE for scholars in the humanities that includes semi-automatic state-of-the-art methods from computer science. This undertaking requires an interdisciplinary approach and intensive collaboration between scholars in the humanities and computer science. On the one hand, DivaDesk should be appropriate for

[1] This stands for Document, Image, and Voice Analysis.

[2] See http://www.e-codices.unifr.ch.

[3] See http://www.manuscripta-mediaevalia.de.

[4] Libraries: presentation of digitized manuscripts; research: edition tools, annotation tools; computer science: automatic DIA methods.

scholars' work, therefore the conception of its generic architecture and functions belongs to the domain of the humanities. On the other hand, it builds on current evolution in the field of computational DIA and is also meant to be the prototypical area of application for ongoing computer science research in that field. The realization and application of DivaDesk, therefore, promise fruitful research outcomes both in the humanities and computer science. The structure of the present paper reflects this interdisciplinary approach. It contains specific information and deals with specific problems in the humanities as well as in computer science in order to address readers in both research communities, while also striving to make the specifics comprehensible for readers in either community.

The rest of this paper is organized as follows. Section 3 summarizes existing tools and projects for historical document viewing and analysis as well as the state of the art of computational methods. Section 4 discusses the current situation in the humanities and conceptualizes the virtual workspace for scholars in the humanities. Existing tools and services from our group are presented in section 5, where we also evaluate several automated DIA tasks. Finally, section 6 concludes the paper and highlights the challenges we are currently facing in interdisciplinary research in the digital humanities.

2. State of the art

Historical document image analysis is considered one of the hardest tasks in the digital humanities for several reasons. For one thing, there are many research questions in the humanities related to individual documents as well as extensive corpora and historical collections. For another, the properties of the physical documents pose many research challenges to automatic DIA. Past research on historical DIA can be categorized in two main tiers: tools developed at institutes in the humanities, typically tailored to the needs of a specific use case, and automatic tools developed by computer scientists in the context of Pattern Recognition (PR). While the border between these categories is not clearly defined and several projects have been conducted jointly by scholars in the humanities and PR researchers (see section 3.1), there is a general tendency toward two divergent movements, as was observed at a recent Dagstuhl Seminar on Digital Humanities.[5]

In the following, we will first review select tools and projects that make use of computational methods in the context of humanity research and, second, summarize important outcomes of PR research in automatic DIA. Finally, toolkits for Ground Truth (GT) generation will be discussed, as GT is an essential requirement for the development of automated DIA methods.

2.1 Historical document viewing and analysis

Many virtual manuscript libraries with different mandates and different functionalities exist. For example, the virtual library called *e-codices* assembles medieval manuscripts in Swiss repositories. It provides a viewer, allows nuanced searches, and enables annotations to be made to specific digitized manuscripts. Close collaboration with libraries ensures sustainable availability of the manuscript images (under a Creative Commons license). Active development to enhance the user interface makes *e-codices* a highly renowned and extensively used virtual library. Other online libraries, such as the *Bibliothèque virtuelle des manuscrits médiévaux*[6] and *manuscripta mediaevalia*, offer similar functionality. Several libraries provide their manuscript collections in useful digital viewers of their own, e.g., the British Library (London),[7] the Bibliothèque nationale de France (Paris),[8] the Houghton Library at Harvard University,[9] the University Library of Heidelberg,[10] and the Herzog August Bibliothek in Wolfenbüttel.[11] One example of a thematically oriented library is the *New Testament Virtual Manuscript Room*,[12] which allows manuscripts to be viewed, indexed, and transcribed. These libraries do not include any automated methods for layout analysis or OCR (Optical character recognition), however.

Tools and projects targeting the transcription of texts appearing in document images have been implemented for specific text corpora. The *Transcribe Bentham*[13] initiative is

[5] Hassner et al. 2013.

[6] http://bvmm.irht.cnrs.fr/.

[7] http://www.bl.uk/manuscripts/.

[8] http://gallica.bnf.fr/.

[9] http://hcl.harvard.edu/libraries/houghton/collections/early_manuscripts/.

[10] http://www.ub.uni-heidelberg.de/helios/digi/handschriften.html.

[11] http://dbs.hab.de/mss/?list=browse&id=project.

[12] http://ntvmr.uni-muenster.de/home.

a collaborative transcription initiative for manuscripts by the philosopher Bentham that relies on a novel crowdsourcing strategy. *Die Rätoromanische Chrestomathie*,[14] another crowdsourcing approach for printed books from the 19th and 20th century, was recently finalized. While the former project does not take advantage of automated recognition by any means, the latter used OCR as initial seed. It has been shown that a recognition error rate lower than 15% is already enough to significantly speed up the transcription process[15] given an appropriate transcription tool when compared with manual transcription without any assistance.

While the tools and projects mentioned above only target the generation of transcriptions, the EU project known as *tranScriptorium*[16] goes one step further. In addition to utilizing DIA results for crowdsourcing applications (e.g., *Transcribe Bentham*), the results are intended to be included in digital archives and e-research portals.[17] The specific objective of *tranScriptorium* is the integration of automated methods into platforms for DIA, text recognition, and keyword spotting. Comparison and presentation aspects are beyond the scope of *tranScriptorium*, however. In the earlier project, IMPACT (*IMProving Access to Text*),[18] a set of automatic tools for specific tasks was developed, which will be reviewed in sections 3.2 and 3.3.

Another initiative targeting automated support is the *Genizah project*,[19] which deals with fragments of Jewish manuscripts collected in a digital library (the *Genizah platform*). An automated search for similar fragments can be performed using local features description to assemble fragments belonging to the same manuscript. As this project deals with the specific case of fragments, OCR and semantic information retrieval functionalities are not included. The ORIFLAMMS project[20] aims at a computer-based paleographic analysis of single characters or parts of characters. Its ultimate goal is to develop an ontology of characters and graphs with the help of automated clustering. The *Monk* system[21] is a platform for word searches in archive collections. It allows for storage and annotation of scanned page images and provides automated methods for text recognition, which deal with problems such as different writing styles for specific words in order to perform search queries on the archival material.

Many tools have been developed for digital editions in recent decades.[22] In the field of Middle High German literature, the *Parzival Project*[23] is one of the most ambitious digital edition projects, dealing with a voluminous text that is transmitted in different textual versions and in more than 80 manuscripts. The *Parzival Project* provides a synoptic edition of different text versions including the possibility of viewing the manuscript image next to the transcribed text as well as a critical edition and digital editions of individual manuscripts on CD or DVD. An interesting feature of this project is the use of the (originally biological) concept of phylogeny to determine and visualize interrelationships between manuscripts.[24] Another powerful tool dealing with textual criticism, phylogeny, and automated collation is CollateX.[25] This tool covers modern machine-readable texts, but does not deal with digitized images.

For digital representation of annotations and comments on document images, the *Shared Canvas platform*[26] has become very popular. It is employed for several use cases (e.g., the Archimedes Palimpsest[27]) and as the underlying architecture for virtual libraries, such as *e-codices*. With well-organized web interfaces, users can easily annotate images and select individual layers for the presentation of the original page and its annotations. The *System for Annotation and Linkage in Arts and Humanities* (SALSAH)[28] provides a VRE for

[13] Causer et al. 2012.

[14] Neuefeind et al. 2011.

[15] Vilar et al. 2010.

[16] Gatos et al. 2014.

[17] http://transcriptorium.eu.

[18] http://www.impact-project.eu.

[19] Wolf et al. 2011.

[20] http://www.irht.cnrs.fr/fr/recherche/les-programmes-de-recherche/oriflamms. See also Stutzmann 2013.

[21] http://www.ai.rug.nl/~lambert/Monk-collections-english.html.

[22] For a survey of the development of digital editions and their theorization, refer to Robinson 2013.

[23] http://www.parzival.unibe.ch/englishpresentation.html.

[24] Viehhauser and Chlench.

[25] http://collatex.net/.

[26] Sanderson et al. 2011.

[27] http://www.archimedespalimpsest.org/.

[28] Schweizer and Rosenthaler 2011.

working with digitized material of all kinds, e.g., manuscripts, books, videos, and even tape recordings. *Shared Canvas* and SALSAH are powerful tools for working with research data and meta-data, but computational DIA methods and automated recognition processes are not included.

Note that this review of tools and projects is far from exhaustive; for a more complete review of other projects and initiatives, such as ENRICH,[29] *manuscriptorium*,[30] and TEI, the reader should refer to Ciocoiu 2012. For a position paper raising general issues in the digital humanities, but focusing on computational methods for paleography, refer to the recent report by the Dagstuhl seminar.[31]

2.2 Computational document image analysis

The majority of DIA methods use binarization, i.e. they separate a given color or greyscale input image into its foreground and background (more than 90% of the recent DIA methods published at ICDAR and ICPR apply binarization at some point). It is apparent that this procedure leads to information loss, while errors made at this stage are inherited in subsequent automated processing steps. It is noteworthy that several recent approaches work without any binarization,[32] but while many text extraction or layout analysis methods require binarized input, the original input image (which should always be kept[33]) can be used again for further processing to ensure that no information is lost.

A set of heuristic methods have been proposed in the literature for the task of binarization, but typically focus on a certain class of documents. Global methods,[34] for example, are extremely fast. Local methods,[35] on the other hand, use different threshold values adaptively deduced for different image regions based on local information. They are strongly dependent on image resolution, the signal-to-background ratio, and local context, thereby making the window size a crucial parameter for a specific document set. Hybrid binarization methods are a straightforward extension of the above methods. They consider both local and global information in the binarization process. Apart from heuristic methods for binarization, machine-learning-based methods also exist. These are applied in two different ways: either (i) by automatically learning the parameters of a given binarization method[36] or (ii) by dividing the image into different regions and learning to select the appropriate method for each region.[37]

Surveys of document layout analysis are found in Mao et al. 2003 and Baird et al. 2011. Document layout analysis is performed in two stages, which are referred to as physical and logical layout analysis respectively. In physical layout analysis, the document image is divided into homogeneous regions depending on their content, e.g. text, graphics, and background. In the succeeding logical layout analysis, these regions are then assigned a specific label, e.g. 'title', 'heading', or 'main text'. A performance analysis of several page segmentation algorithms is presented by Shafait et al. 2008. Text-line detection methods are typically performed with an error rate of around 5%.

Research on text recognition (also often referred to as OCR for printed text, and Handwriting Recognition (HWR) for handwritten text) has been carried out for more than five decades.[38] The current state of the art is performing recognition using Recurrent Neural Networks (RNN) with error rates as low as <1% for printed historical documents and 6% for medieval manuscripts.[39] For more information on these methods, refer to section 5.3.

2.3 Ground Truth generation tools

An indispensable prerequisite for developing reliable semi-automated methods is to incorporate experts' knowledge in such systems ('learn from the expert'). This knowledge needs to be provided in the form of correct labels for the information extractable from digital facsimiles, called 'Ground Truth' or 'GT' in DIA research. GT facilitates the development of robust automated DIA approaches by enabling a machine to learn by example and, furthermore, allows assessing its performance by referring to the correct labels, i.e. how close

[29] http://enrich.manuscriptorium.com/.

[30] http://www.manuscriptorium.com/.

[31] Hassner et al. 2013.

[32] Chen et al. 2014; Garz et al. 2012.

[33] E.g., by building on the International Image Interoperability Framework (IIIF); see http://iiif.io.

[34] Otsu 1975; Yang et al. 2006.

[35] Niblack 1990; Trier, and Jain 1995; Sauvola and Pietikäinen 2000.

[36] Chou et al. 2010; Chamchong and Fung 2010.

[37] Sari et al. 2012.

[38] Plamondon et al. 2000; Bunke 2003.

[39] Fischer et al. 2012.

the prediction of the system is to the real data as labeled by a human expert. Hence, the primary aim of our project is to facilitate fast and uncomplicated generation of GT for large amounts of digitized documents.

Several labeling tools have been developed in recent years. AGORA[40] segments (historical) document images into two maps in order to divide them into the fore- and background. A user can then label, merge, and remove computed regions. *PixLabeler*[41] is a pixel-level labeling tool for binarized (bi-tonal) document images, where foreground pixels are assigned a color, with each color representing a label such as 'handwriting', 'machine print', 'graphics', etc. *Aletheia*[42] follows a top-down approach (iteratively splitting regions into smaller entities) for labeling binarized document images. Regions automatically detected by splitting and shrinking (fitting the boundary of a region to the entity) can be modified by the user, while low-level elements can be aggregated into more complex structures.

3. DivaDesk, a digital workspace for analyzing historical documents

3.1 Discussion of the current situation

As highlighted in the introduction, technical developments have had a considerable impact on the humanities. Working with historical documents has been simplified by the availability of digitized manuscripts and prints, while the way of working has also changed dramatically. Digitized manuscripts are more easily and readily accessible than microfilm or original documents, but this facilitated access also has its drawbacks. Obviously, the impression of the whole physical object is not available in a digital viewer and specific properties, e.g., binding and watermarks, can only be investigated on the original document. More crucially, the presentation of digitized material leads to a change in the behavior of researchers when investigating the material. Instead of considering the manuscript as a whole, in most cases they selectively access only a few pages, parts of texts, and other points of interest. As a result, relations the viewed page has to other pages or parts of the manuscript or general properties of the whole manuscript remain unnoticed. During an examination of the physical object, such discoveries are often made unintentionally or by chance and lead to novel insights. A digital tool for scholars in the humanities should, therefore, take the aforementioned workflow into account and facilitate fast investigation of whole manuscripts by including functionalities to automatically identify potential regions of interest and notify the scholar, while equally allowing for new relationships the system did not come up with.

Another obstacle is the cumbersome way of accessing documents through diverging interfaces and function modes of different viewing tools employed by libraries or manuscript databases. If a scholar works on manuscripts held in different repositories, he has to adapt his working procedure to the specifics of different viewing tools; this reinforces the tendency to mainly perform specific, short-sighted searches in order not to lose too much time in adapting to unknown interfaces.

It is therefore crucial to build a digital workspace for scholars in the humanities that meets professional standards and allows intuitive handling of digitized data at the same time. The conceptualization of such a workspace is the main purpose of this article. The workspace should support the scholar in his daily work and in the digital presentation of research outcomes by making use of computational methods without forcing him to adapt his working procedures to implementation-specific requirements that different tools have for specific tasks.

3.2 Interdisciplinary approach

The aim of our work is to create a workspace that allows scholars to perform research in their accustomed way while being supported by state-of-the-art computational DIA methods. Therefore, an interdisciplinary approach is of crucial importance. Computer scientists and scholars in the humanities will collaborate intensively during the entire development and evaluation process of the tool. This will enable relevant functionality and good usability to be provided for the tool and thus ensure its acceptance by the research community.

One example of the interdisciplinary approach we are taking is the projected workflow for (semi-)automatic script analysis (for further information on the technical issues of this task, see Garz et al. 2014). After the development of a prototype framework by computer scientists, scholars in the humanities will assist them in finding appropriate test data (manuscript images) along with corresponding GT. Initial experiments will then be performed on this data, and the results of these will then be reviewed and evaluated by scholars in the humanities by means of GT and feedback

[40] Ramel et al. 2006.

[41] Saund et al. 2009.

[42] Clausner et al. 2011.

on the result representation. This iteratively performed dual evaluation is a very important step in the improvement of the tool.

3.3 Conception

In this section, we present our conception of DivaDesk, which addresses specific research with historical documents. It is noteworthy that these concepts can be transferred to other areas in the humanities as well. In the VRE, the scholar is able to gather digital facsimiles of manuscripts he is working on and arrange them in virtual libraries in order to reconstruct historical collections (i.e., manuscripts having once belonged to the same library, but now dispersed in different repositories) or to assign a manuscript to one or more of his research topics, such as different testimonials of the same text for an edition project. The scholar can describe the manuscripts or import existing catalogs, transcribe texts or parts of the text from the primary source, and make local annotations for specific spots inside a manuscript or extensive annotations on a more general level. All these tasks are not fundamentally different from the usual procedure of scholars working with historical documents, as they generate and store their research outcomes as text files on a computer or as paper files in a physical folder. The scholar is therefore not forced to adapt his procedure to a standard interface developed according to common practice in computer science, but the virtual workspace offers essential advantages with respect to usability, connection, and presentation of the research outcomes when compared with a real writing desk.

The fact that the whole working process – from the description of the manuscript to the finished article on the topic – can take place in the same VRE and that outcomes of previous working processes are also accessible in this environment allows for synergistic effects. By means of a (semi-)automated or manual search, similarities or relations between different documents or research outcomes can easily be discovered or retrieved. Let us illustrate this with the example of automated watermark retrieval, which we plan to integrate into DivaDesk. If scholars work on an inventory of several dozen manuscripts, they are very unlikely to remember the watermarks of the first manuscript when describing the 30th manuscript. Yet a relation between two manuscripts could be discovered through an identical watermark. An automatic image-retrieval process can point out similarities between watermarks in different manuscripts and hence help scholars to cope with large collections of research material. Furthermore, DivaDesk will support scholars in visualizing and presenting research outcomes. The joint presentation of research outcomes in textual and visual form, e.g., transcribed text together with an image of the original document and visualizations of relations between manuscripts or contents, is particularly useful for research inspired by the theory of New Philology[43] or guided by the interest in the materiality of the documents.[44]

3.4. Architecture

The DivaDesk workspace currently focuses on the analysis of medieval manuscripts, but its modular architecture allows for the integration of services for other kinds of documents, e.g., prints, (early) modern handwritten documents, pictures, and objects. Its specification allows for the integration of particular services or sets of services into other frameworks as well as the integration of external services into DivaDesk using specified protocols. The interoperability of DivaDesk with other frameworks and platforms is thus guaranteed (see section 5 for technical details).

The architecture of the workspace for medieval manuscripts builds on three main modules, as depicted in fig. 1, and contains two different interaction modes. The input mode supports the working tasks of the scholar, e.g., cataloging, transcribing, and arranging digitized data. The presentation mode allows data to be displayed, either as an intermediate stage of the working process or as a research outcome.

The first module addresses the description of items — medieval manuscripts in our case. It contains information about each item in the form of a catalog. In accordance with the *TEI-P5 format for manuscript description*,[45] it is composed of three submodules:

1. the physical description of the manuscript;
2. the history of the manuscript containing its origin and provenance;
3. an overview of the content.

This information is generated by the scholar while working on the manuscript or is imported from existing manuscript catalogs. In the submodule *Physical Description*, several (semi-)automated processes can be used: layout analysis for

[43] Gleßgen and Lebsanft 1997.

[44] Nichols et al. 1996; Nichols 1997; Ortlieb 2013.

[45] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/MS.html.

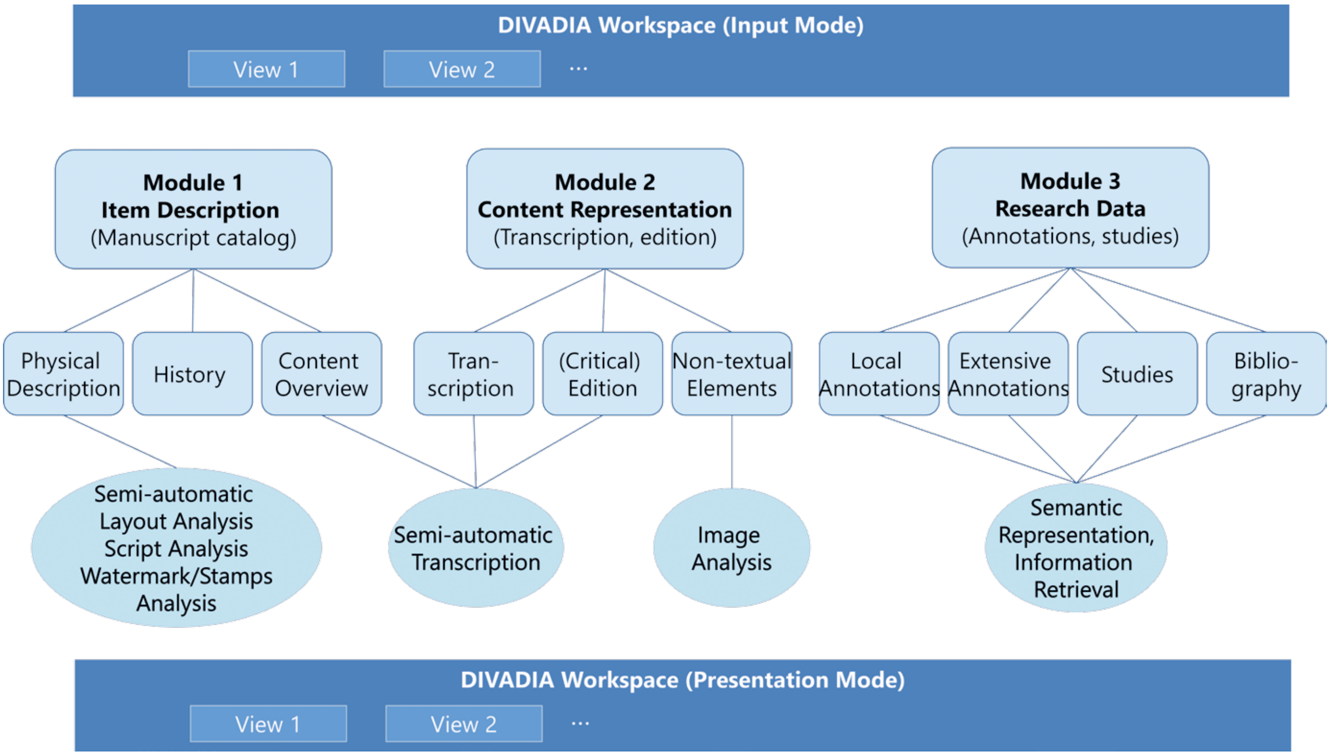

Fig. 1: Overview of DivaDesk's architecture.

structuring the pages and finding points of interest; script analysis for identifying scribe changes; and watermark and binding stamp analysis for identifying and comparing similar watermarks/stamps.

The second module is concerned with the representation of the content. For medieval manuscripts, it contains three submodules:

1. a submodule for transcription built according to the *TEI-P5 Format Representation of Primary Sources*[46] and linked to the digital facsimile;
2. a submodule that allows for (critical) editions of several textual witnesses/versions and is built on the *TEI-P5 format Critical Apparatus*;[47]
3. a submodule for the representation of non-textual elements, i.e., decorations, miniatures, or diagrams. (Semi-)automatic text-recognition processes will be used for the transcription.

The third module contains research data created by the scholar that is neither a description/catalog of the manuscript nor a representation (transcription/edition) of the primary source. In this way, we are building on the foundations laid out in DARIAH.[48] The four submodules of this module allow for different ways of linking the research data to the digitized data:

1. local annotations are linked to a specific spot in the digital facsimile, e.g., a word, text block or image;
2. extensive annotations are not bound to a specific local area within the facsimile having a larger/more abstract scope, e.g., summaries, plot patterns, or motifs;
3. in contrast to the annotation submodules, the third submodule contains finished (and/or published) studies discussing one or several manuscripts that are part of the virtual workspace. For the semantic annotation, XML-based methods and tools are already available, e.g., the project *Sharing Ancient Wisdoms*,[49] which can be integrated and adapted in the DivaDesk workspace;
4. a bibliography of secondary literature mentioning manuscripts that are part of the virtual workspace. In this module, information retrieval methods will be used, e.g., for finding shelfmark mentions in secondary

[46] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/PH.html.

[47] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TC.html.

[48] http://dariah.eu/.

[49] Jordanous et al. 2012.

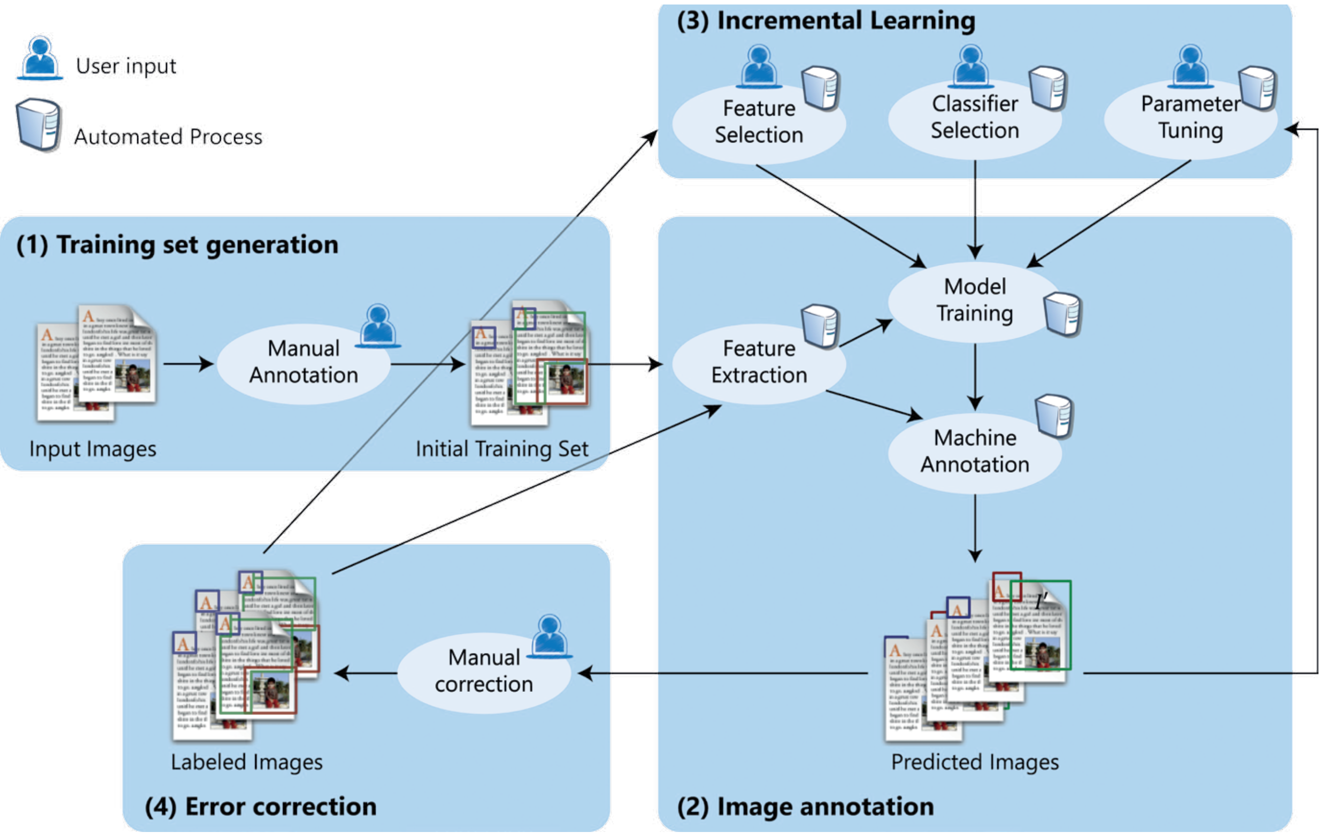

Fig. 2: Workflow of the document-labeling process.

literature and linking this information to the respective facsimiles in the workspace.

4. Technical realization and implemented tools & services

DivaDesk is a modular software system implementing several state-of-the-art methods for the modules illustrated in fig. 1 as well as views for input and presentation. The system can easily be expanded by plugging in new modules offering additional functionality. Currently, a set of services called DivaDia Services has already been realized. These services are used by the DivaDia labeling tool.[50] While DivaDia is based on Java, any other programming language can be used; for *Item Description and Transcription*, for example, DivaDiaWT, an HTML5 view (based on the AngularJS framework), has been implemented (see section 5.4).

The modular design of our system allows for seamless integration of existing tools for specific tasks into the overall framework as well as the integration of components into other platforms. The computational methods generated during the IMPACT and tranScriptorium project, for example, enhance the image-processing functionality of Module 1. Furthermore, the open-source OCR engine OCROPUS[51] provides state-of-the-art layout analysis and text recognition. Since we are building on the Shared Canvas framework and use TEI for the representation of document meta-data, data created in our tools can be shared with other platforms, such as *e-codices*. Our target-oriented web service (e.g., text-line segmentation) can be used by other transcription tools, and even the web-based user interface for the transcription of documents can be integrated into other online platforms, such as the new version of *e-codices* and SALSAH. Such interaction with other projects allows for a holistic framework that serves all aspects of scholarly work with historical manuscripts and research data, incorporating the recent outcomes of research in automatic DIA.

While at present the holistic workspace is still in the conceptualization phase, several web services have already been implemented as DivaDia Services, which will be integrated

[50] A preliminary Java version is available at http://diuf.unifr.ch/hisdoc/divadia. The Web-Interface and RESTful Webservices are only accessible inside the local network of the University of Fribourg, but will be opened as soon as the administrative process at our University has been completed. The web services have preconfigured settings for common workflows, such as image and text retrieval in the facsimiles.

[51] http://code.google.com/p/ocropus/.

into the three modules of the DivaDesk architecture. In this section, we describe the functionality of these services. Subsequently, we provide an overview of existing views/interfaces for specific workflows. Finally, we conclude with evaluation results on publicly available data sets and an outlook on future work.

4.1 Module 1: Item Description

The main concept used in this module is incremental learning, i.e. the system adjusts what it has learned previously according to new examples. We illustrate this with an example: Given that an automated DIA method works well on one set of historical documents, the recognition performance might still not be optimal for a specific unseen manuscript. In such a case, the user labels a few samples (text lines or pages). These labels are then used as GT for adapting and improving the automated DIA methods. The detailed process of incremental learning is described in the following and illustrated in fig. 2.

In order to start a new labeling process, a user manually annotates several[52] representative lines or page images of a document (1). These images in conjunction with their annotations compile the GT for the generation of an adapted prediction model, which is computed on the basis of a feature set and a machine-learning (ML) algorithm (2). Having computed the adapted model, the user tests it by selecting another set of images (3), which is automatically annotated based on the model. The predicted result is presented to the user, who can manually correct and/or accept it (4a) or try to improve the model (4b) by changing the ML algorithm, feature set, or parameters, for instance. If the user accepts a result, the GT is extended (5) by those newly labeled images, and a refined prediction model is computed. Starting from the third step, the process is pursued until the entire document has been annotated.

In DivaDia Services, the following DIA methods are currently available:

- Image processing: standard binarization methods, local filters such as edge detectors, smoothing, Laplacian of Gaussian (LoG), Difference of Gaussians (DoG), and other non-linear filters that help enhance the image and remove noise.
- Feature extraction:[53] color and coordinates, incorporating information on the neighborhood of the considered pixel, Local Binary Patterns (LBP) that focus more on the textual structure of the image, Scale-Invariant Feature Transform (SIFT), which considers 'interesting' regions in the image, and Gabor features, which describe the dominant orientation of a pixel. Currently, unsupervised feature-learning methods are implemented as well.
- Feature selection: greedy forward/backward selection, sequentially floating forward selection, linear forward selection, genetic selection, and hybrid feature selection.[54] Feature selection methods help to retrieve the best set of features by automatically testing several combinations and systematically searching for the best combination.
- Machine-learning algorithms that automatically learn to classify given patterns after receiving sample patterns with GT: Support Vector Machines (SVMs), Modified Quadratic Discriminant Function (MQDF), k-nearest neighbor algorithm (k-NN), Naïve Bayes classifier (NB), Gaussian Mixture Models (GMMs), Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM),[55] and Markov Random Fields (MRF).
- Evaluation: Presently, our evaluation metrics are based on precision and recall. Additional metrics will be incorporated to enable a user to assess the quality of the prediction results. The annotation information is saved in XML format.[56]

These methods and all future versions of DivaDia Services are available as open-source projects. Note that the methods are generally able to deal with texts written in any scripts and orienting any direction, but the trained methods only work on Latin script. For a description of technical details, refer to Wei et al. 2013.

4.2 Module 2: Content Representation

The concept of the automated DIA methods for content representation is similar to the one described for Module 1. Furthermore, most of the approaches described above can be used for recognition and analysis of non-textual

[52] Note that the exact number of text lines or page images to be manually annotated depends on the specific difficulties of a given manuscript.

[53] Wei et al. 2013.

[54] Wei et al. 2014.

[55] Graves et al. 2009.

[56] Pletschacher and Antonacopoulos 2010.

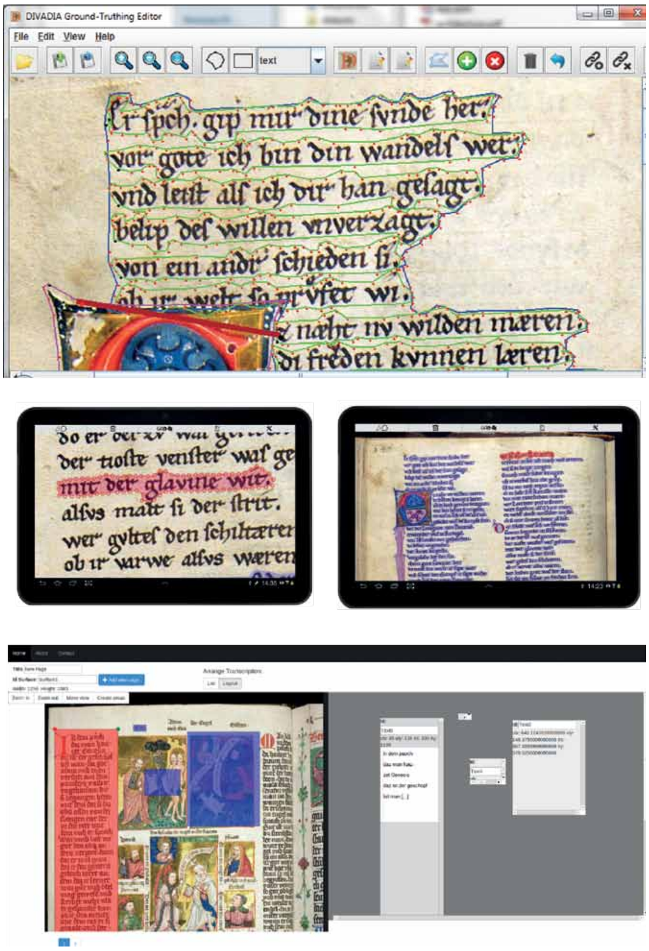

Fig. 3: DivaDia labeling tool (top row), Android application (second row) and DivaDiaWT (bottom row).

images as well, especially SIFT and the ML methods. We have developed a set of methods in the HisDoc project for recognizing texts (OCR/HWR) and for editing.[57] These methods rely on handwritten text lines as input. After normalization of skew (inclination of the text line), slant (inclination of the characters), and width and height, one out of two state-of-the-art text-line recognition modules can be applied: Hidden Markov Models (HMM)[58] and LSTM[59] neural networks. The results of these methods as well as the manually entered information are stored in TEI format in order to ensure compatibility with other existing and future toolkits and frameworks.

4.3 Module 3: Research Data

Currently, our services allow researchers to find specific text passages in manuscripts even if the automatic recognition did not perform perfectly or the query is ambiguous, e.g., due to orthographic or dialectal varieties, similar to Google search. In order to support generation and access research data, we will integrate state-of-the-art Natural Language Processing (NLP) and Information Retrieval (IR) methods in the future.[60] Furthermore, we plan to integrate the identification of manuscript shelfmarks in order to automatically link existing annotations or published studies concerning the same manuscripts.

4.4 Views

The concept of DivaDesk incorporates several views (or interfaces) for the input and (re-)presentation of item descriptions, content, and research data. Note that most of the views can be used for input as well as for presentation of the data, depending on current needs, thereby reducing the learning curve for a new user. For example, a view for semi-automatic transcription of manuscript pages can be used for viewing the transcriptions as well. In the following, three existing prototypical views are presented.

The DivaDia labeling interface allows users to manually or (semi-)automatically label a document image. They can display and enhance an image using several image-manipulation methods in order to make details visible, for example. Drawing tools similar to those known from image-editing software (e.g., Adobe Photoshop, Gimp) allow manual annotation of regions. In order to modify the automatic prediction of the system, a user can change system parameters after visually inspecting the previewed results or directly modify the size, boundaries, position, or category of predicted regions. A screen shot of this interface in its current form is presented above in fig. 3. In addition, we have implemented an interface for mobile Android devices supporting *touch & write* input[61] (see the second row in fig. 3), providing a natural user interface for scholars in the humanities, as annotations can be directly drawn and written with a pen. Finally, DivaDiaWT is a web-based interface that presents the transcriptions in the layout of the original manuscript image (see bottom row in fig. 3). It includes

[57] Fischer et al. 2012.

[58] Ploetz and Fink 2009.

[59] Graves et al. 2009.

[60] Naji and Savoy 2011.

[61] Liwicki et al. 2010; Dengel et al. 2012.

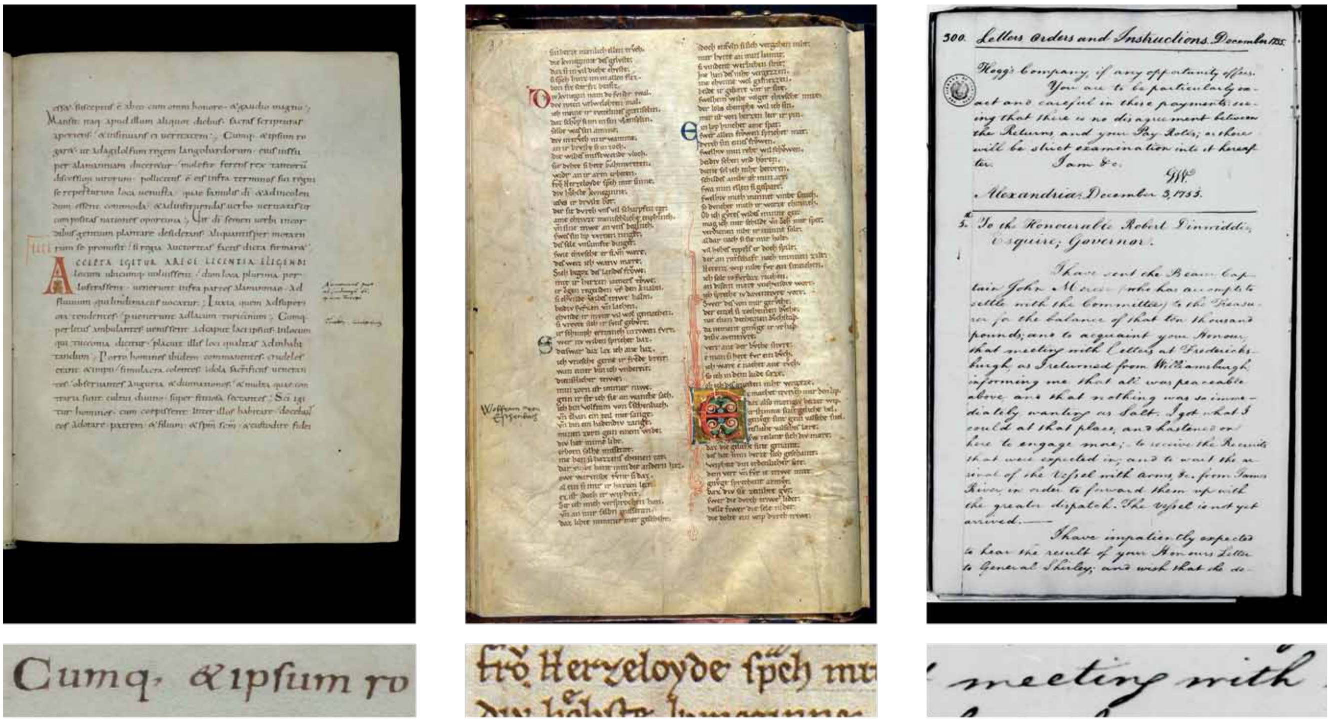

Fig. 4: From left to right: exemplary images of the Saint Gall data set (St. Gallen, Stiftsbibliothek, Cod. Sang. 562, p. 7), the Parzival data set (St. Gallen, Stiftsbibliothek, Cod. 857, p. 36), and the George Washington data set (Washington, Library of Congress, Letterbook 1, p. 300).

automatic support for line segmentation and the generation of XML files in TEI without having to install any software.

4.5 Evaluation of automated methods

The methods developed throughout the completed HisDoc project,[62] the ongoing project HisDoc 2.0,[63] and methods implemented in the DivaDia Services (see sections 5.1 and 5.2) have been constantly evaluated on data sets of a diverse nature. More specifically, the three data sets of the IAM-HistDB[64] have been used for evaluation: the Saint Gall data set,[65] the Parzival data set,[66] and the George Washington data set.[67] For sample images, refer to fig. 4. Note that for purposes of brevity, most technical details (exact sizes of training and test set, algorithm- and data-specific parameters, and detailed analyses) do not appear in this article, but are available in the cited publications. In this article, we focus on the most significant outcomes of these experiments.

The performance of our layout analysis methods is measured at pixel level. When considering a four-class classification problem, i.e. categorizing the pixels into periphery, background, text block, and decoration, our best method achieves an error rate of less than 9% on the Parzival data, around 4% on the Saint Gall data, and 11% on the George Washington data[68] with the Naïve Bayes ML algorithm, which runs very fast. These results are already useful for practical applications, since errors only appear on the border of the lines and, typically, a perfect border is not required for recognition.

For text recognition, i.e. OCR on complete text lines, we achieved word-level error rates of around 3.5% on the two medieval data sets when text lines were segmented manually. When applying our fully automated system, which first detects text lines and then recognizes the text, the error rate increases to 7%,[69] but even this result is useful in practice. As mentioned in section 3.1, integrating an automatic system with an error rate of less than 15% into the transcription

[62] Fischer et al. 2012.

[63] Garz et al. 2014.

[64] http://www.iam.unibe.ch/fki/databases/iam-historical-document-database.

[65] St. Gallen, Stiftsbibliothek, Cod. Sang. 562.

[66] St. Gallen, Stiftsbibliothek, Cod. Sang. 857.

[67] Washington, Library of Congress, Letterbook 1.

[68] Wei et al. 2014.

[69] Fischer et al. 2014.

process would speed up the annotation process significantly. To assess the performance of Information Retrieval (IR), we simulated 60 possible user queries searching for text appearances in manuscripts of the Parzival data set, e.g., ‘dem man dirre aventivre giht’, ‘iwer oder decheines man’, and ‘als man von siner helfe saget’. System performance is measured by the mean reciprocal rank, a measure that is high if the text lines of interest appear in the top ranks. We have observed that the performance loss of IR on automatically recognized text is less than 1% compared with the performance on perfect transcription.[70] This result confirms that the OCR is already useful in practice.

4.7 Future work

Apart from the technical realization of DivaDesk, the quality of source material and legal issues are going to be major challenges in future. In order to achieve high-recognition performance, the quality of the digitized images needs to be sufficiently high (300 dpi is considered to be the minimum resolution for our current methods, and lossless image formats are indispensable). We are currently studying methods that work on low-resolution images in order to integrate images from other sources than recent high-quality digitization, e.g., digitized microfilm or historical photographs of lost or destroyed manuscripts.

Legal issues include copyright on image data, which typically remains with the repositories of the original documents. In the present development phase, we have not started focusing on those issues yet, given that our primary aim is to build an individual workspace for personal research data. As long as the data is used in the individual workspace only, legal issues are no major problem. We are aware of the importance of these issues for later stages of the realization of our workspace when functionalities for publishing and sharing research data will be added. Those issues can only be solved by all the stakeholders concerned cooperating with each other.

The finished HisDoc project and the ongoing project, HisDoc2.0, focus on the development of computational methods for DIA and OCR mainly from the computer scientist’s point of view. The development of usable tools and a workspace for scholars in the humanities is a long-term goal to which these projects will lead; DivaDesk’s conception and design are fine for the current project phase, but its realization is not included in the planned outcomes of the HisDoc 2.0 project. In order to realize a functional open-access version of the DivaDesk VRE, we envisage follow-up projects with an interdisciplinary focus.

5. Conclusion and outlook

In this paper, we proposed a novel conception of a holistic digital workspace called DivaDesk. It allows scholars working with historical documents to continuously make full use of new possibilities arising from technological development and digitization. The workspace provides a set of services that are interoperable with other frameworks and platforms. Therefore, individual combinations of different VREs are possible. The architecture of DivaDesk consists of three main modules: Item Description, Content Representation, and Research Data. It provides several state-of-the-art computational methods for supporting scholars in the humanities in their daily work. Current implementations provide automated DIA methods that achieve cutting-edge performance for layout analysis and text recognition. The envisioned workspace will provide further functionalities, supporting high-performance searches, comparison of editions and texts, and seamless connections to diverse research data. DivaDesk will be useful for research in the humanities and will push the current limits of the art in DIA.

[70] Fischer et al. 2012.

REFERENCES

Baird, H. S., Bunke, H., and Kazuhiko, Y. (1992/2011), *Structured Document Image Analysis* (1st ed.), (Berlin – Heidelberg: Springer), (DOI: 10.1007/978-3-642-77281-8).

Bunke, H. (2003), 'Recognition of cursive Roman handwriting: past, present and future', *Seventh International Conference on Document Analysis and Recognition. Proceedings*, 448-459.

Causer, T., Tonra, J., and Wallace, V. (2012), 'Transcription maximized; expense minimized? crowdsourcing and editing *The Collected Works of Jeremy Bentham'*, *Literary and Linguistic Computing,* 27.2: 119–137.

Chamchong, R., and Fung, C. (2010), 'Optimal selection of binarization techniques for the processing of ancient palm leaf manuscripts', in: *Systems Man and Cybernetics (SMC)*: 3796–3800.

Chen, K., Wei, H., Liwicki, M., Hennebert, J., and Ingold, R. (2014), 'Robust Text Line Segmentation for Historical Manuscript Images using Color and Texture', in *22nd International Conference on Pattern Recognition*, 2978-2983.

Chou, C.-H., Lin, W.-H., and Chang, F. (2010), 'A binarization method with learning-built rules for document images produced by cameras', *Pattern Recognition*, 43.4: 1518–1530.

Ciocoiu, A. M. (2012), *International collaboration in digital libraries: an analysis of the Manuscriptorium digital library case study*, Master thesis: International Master in Digital Library Learning, University of Tallinn/University of Parma (http://hdl.handle.net/10642/1266 [last accessed 13 Jan. 2015]).

Clausner, C., Pletschacher, S., Antonacopoulos, A. (2011), 'Aletheia — An Advanced Document Layout and Text Ground-Truthing System for Production Environments', in *International Conference on Document Analysis and Recognition*, 48–52.

Dengel, A., Liwicki, M., and Weber, M. (2012), 'Touch & Write: Penabled Collaborative Intelligence', in *Knowledge Technology,* 1–10 (Berlin – Heidelberg: Springer).

Fischer, A., Bunke, H., Naji, N., Savoy, J., Baechler, M., and Ingold, R. (2012), 'The HisDoc Project: Automatic Analysis, Recognition, and Retrieval of Handwritten Historical Documents for Digital Libraries', in *The Proceedings of InterNational and InterDisciplinary Aspects of Scholarly Editing*.

——, Baechler, M., Garz, A., Liwicki, M., and Ingold, R. (2014), 'A Combined System for Text Line Extraction and Handwriting Recognition in Historical Documents', in *International Workshop on Document Analysis Systems*, 71–75.

Garz, A., Fischer, A., Sablatnig, R., and Bunke, H. (2012), 'Binarization-Free Text Line Segmentation for Historical Documents Based on Interest Point Clustering', in *International Workshop on Document Analysis Systems*, 95–99.

——, Eichenberger, N., Liwicki, M., and Ingold, R. (2014), 'HisDoc 2.0 – Towards Computer-Assisted Paleography', *manuscript studies,* 7.

Gatos, B., Louloudis, G., Causer, T., Grint, K., Romero, V., Sánchez, J. A., Toselli, A. H., and Vidal, E. (2014), 'Ground-truth production in the transcriptorium project', in *International Workshop on Document Analysis Systems*, 237-241.

Gleßgen, M.-D., and Lebsanft, F. (eds.) (1997), *Alte und neue Philologie* (Tübingen: Niemeyer).

Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H., Schmidhuber, J. (2009), 'A novel connectionist system for improved unconstrained handwriting recognition', in *IEEE Transactions on Pattern Analysis and Maschine Intelligence,* 31: 855–868.

Hassner, T., Rehbein, M., Stokes, P. A., and Wolf, L. (2013), 'Computation and Palaeography: Potentials and Limits (Dagstuhl Perspectives Workshop 12382)', *Dagstuhl Reports*, 2.9: 14-35.

Jordanous, A., Lawrence, K. F., Hedges, M., and Tupman, C. (2012), 'Exploring manuscripts: sharing ancient wisdoms across the semantic web', in *International Conference on Web Intelligence, Mining and Semantics*, 44.

Liwicki, M., Rostanin, O., El-Neklawy, S. M., and Dengel, A. (2010), 'Touch & write: a multi-touch table with pen-input', in *International Workshop on Document Analysis Systems*, 479–484.

Mao, S., Rosenfeld, A., and Kanungo, T. (2003), 'Document structure analysis algorithms: a literature survey', in *Electronic Imaging* (International Society for Optics and Photonics), 197–207.

Naji, N., Savoy, J. (2011), 'Information retrieval strategies for digitized handwritten medieval documents', in *Asia Conference on Information Retrieval Technology*, 103–114.

Neuefeind, C., Rolshoven, J., and Steeg, F. (2011), 'Die Digitale Rätoromanische Chrestomathie – Werkzeuge und Verfahren für die Korpuserstellung durch kollaborative Volltexterschließung', in *Multilingual Resources and Multilingual Applications: Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology*.

Niblack, W. (1990), *An Introduction to Digital Image Processing* (Upton Saddle River, NJ: Prentice Hall , Inc.).

Nichols, S. G. (1997), ‘Why material philology?’, in Tervooren, H., und Wenzel, H. (eds.), *Philologie als Textwissenschaft. Alte und neue Horizonte* (Berlin: Schmidt), 10–30.

——, et al. (eds.) (1996), *The Whole Book. Cultural Perspectives on the Medieval Miscellany* (Ann Arbor: Univ. of Michigan Press).

Ortlieb, Cornelia (2013), ‘Materialität’, in R. Borgards et al. (eds.), *Literatur und Wissen. Ein interdisziplinäres Handbuch* (Stuttgart: Metzler), 41–45.

Otsu, N. (1975), ‘A threshold selection method from gray-level histograms’, in *Automatica*, vol. C, no. 1: 62–66.

Plamondon, R., and Srihari, S. N. (2000), ‘Online and off-line handwriting recognition: a comprehensive survey’, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 22.1: 63–84.

Pletschacher, S. and Antonacopoulos, A. (2010), ‘The PAGE (Page Analysis and Ground-Truth Elements) Format Framework’, in *International Conference on Pattern Recognition*, 257–260.

Ploetz, T., Fink, G.A. (2009), ‘Markov models for offline handwriting recognition: a survey’, *International Journal on Document Analysis and Recognition*, 12: 269–298.

Ramel, J. Y., Busson, S., Demonet, M. L. (2006), ‘AGORA: the Interactive Document Image Analysis Tool of the BVH Project’, in *International Conference on Document Image Analysis for Libraries*, 145–155.

Robinson, P. (2013), ‘Towards a Theory of Digital Editions’, *Variants,* 10: 105–131.

Sanderson, R. Albritton, B. Schwemmer, R. Van de Sompel, H. (2011), ‘SharedCanvas: A Collaborative Model for Medieval Manuscript Layout Dissemination’, *ACM/IEEE Joint Conference on Digital Libraries*, *Ottawa, Canada, June 2011,* 175-184.

Sari, T., Kefali, A., and Bahi, H. (2012), ‘An MLP for binarizing images of old manuscripts’, in *International Conference in Frontiers in Handwriting Recognition*, 247–251.

Saund, E., Lin, J., Sarkar, P. (2009), ‘PixLabeler, User Interface for Pixel-Level Labeling of Elements in Document Images’, *International Conferenceon Document Analysis and Recognition*, 646–650.

Sauvola, J., and Pietikäinen, M. (2000), ‘Adaptive document image binarization’, *Pattern Recognition,* 33.2: 225–236.

Schweizer, T., and Rosenthaler, L. (2011), ‘SALSAH – eine virtuelle Forschungsumgebung für die Geisteswissenschaften’, *Electronic Visualisation and the Arts Berlin* (Elektronische Medien & Kunst, Kultur, Historie. 9.-11. November 2011), 147–153.

Shafait, F., Keysers, D., and Breuel, T. (2008), ‘Performance evaluation and benchmarking of six-page segmentation algorithms’, *Pattern Analysis and Machine Intelligence*, 30.6: 941–954.

Stutzmann, D. (2013), ‘Système graphique et normes sociales: pour une analyse électronique des écritures médiévales’, in N. Golob (ed.), *Medieval Autograph Manuscripts* (Bibliologia, 36), 429–434.

Trier, O., and Jain, A. (1995), ‘Goal-directed evaluation of binarization methods’, in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17.12: 1191–1201.

Viehhauser, G., and Chlench, K. (2014), ‘Phylogenese und Textkritik. Bioinformatische Anregungen zur Lösung genealogischer Klassifizierungsprobleme in der Editionsphilologie’, in M. Stolz (ed.), *Internationalität und Interdisziplinarität der Editionswissenschaft,* (Berlin: de Gruyter; Beihefte zu Editio, 34), 57-82.

Vilar, J. M., Castro-Bleda, M. J., Zamora-Martínez, F., España-Boquera, S., Gordo, A., Llorens, D., Marzal, A., Prat, F., Gorbe, J. (2010), ‘A flexible system for document processing and text transcription’, in *Current Topics in Artificial Intelligence* (Berlin – Heidelberg: Springer), 291–300.

Wei, H., Baechler, M., Slimane, F., Ingold, R. (2013), ‘Evaluation of SVM, MLP and GMM Classifiers for Layout Analysis of Historical Documents’, in *International Conference on Document Analysis and Recognition*, 1252–1256.

——, Chen, K., Ingold, R., and Liwicki, M. (2014), ‘A Hybrid Feature Selection Method for Historical Document Image Analysis’, in *International Conference on Frontiers in Handwriting Recognition*.

Wolf, L., Dershowitz, N., Potikha, L., German, T., Shweka, R., Choueka, Y. (2011), ‘Automatic Palaeographic Exploration of Genizah Manuscripts’, in *Kodikologie und Paläographie im digitalen Zeitalter - Codicology and Palaeography in the Digital Age*, ed. by F. Fischer, C. Fritze, and G. Vogeler (Norderstedt Schriften des Instituts für Dokumentologie und Editorik, 3), 157-179.

Yang, Z. Ma, and M. Xie (2006), ‘A novel binarization approach for license plate’, in *IEEE Conference on Industrial Electronics and Applications*, 1–4.