# Machines and Manuscripts 2015
# 2nd International Workshop on Opportunities for the Automatic Pattern Recognition and Analysis of Historical Documents

## February 19-20, 2015
## Schlossbezirk 13, Bldg 20.13 Raum 001, 76131 Karlsruhe

Digital methods, tools and algorithms are gaining importance for the analysis of the digitized manuscripts collections in the arts and humanities. The goal of this workshop is to bring together experts from diverse fields, such as Pattern Recognition, Digital Humanities, and Information Visualization.

The workshop is organized by the project "eCodicology – Algorithms for the Automatic Tagging of the Medieval Manuscripts, supported by BMBF and hosted by Karlsruhe Institute of Technology (KIT), Germany.

## Program for 19th February 2015

### *Welcome Speech*

13:00-13:05: Opening (*Andrea Rapp)*

13:05-13:20: Computer Science at Karlsruhe Institute of Technology (*Michael Beigl)*

13:20-13:30: Organizational Information (*Rainer Stotzka)*

### *Session I: Project Introduction*

*Moderation: Andrea Rapp*

13:30-13:50: Introduction to eCodicology – Algorithms for the Automatic Tagging of Medieval Manuscripts *(Andrea Rapp, Darmstadt)*

13:50-14:05: The Technical Perspective of eCodicology *(Swati Chandna, Karlsruhe)*

The BMBF funded joint research project "eCodicology" aims to design, evaluate and optimize algorithms for the automatic identification of macro- and micro- structural layout features of the medieval manuscripts.

The main goal of this research project is to provide humanities scholars better insights into high dimensional datasets of medieval manuscripts. The heterogeneous nature of the large humanistic data and the need to create a database of automatically extracted reproducible features for better statistical and visual analysis, are the main challenges in designing a workflow for the arts and humanities. As a solution, this paper presents a concept of a workflow for the automatic tagging of medieval manuscripts. As a starting point, the workflow uses medieval manuscripts digitized within the scope of the project “Virtual Scriptorium St. Matthias". The metadata schema and the models for XML _les are created according to TEI P5. Firstly, these digitized manuscripts are ingested into a data repository. Secondly, specific algorithms are adapted or designed for the identification of macro- and micro-structural layout elements like page size, writing space, number of lines etc. And lastly, the statistical analysis and scientific evaluation of the manuscripts groups are performed. The workflow is designed generically to process large amount of data automatically with any desired algorithm for feature extraction. As a result, a database of objectified, reproducible features is created which helps to analyze and visualize hidden relationships of around 170,000 pages. The workflow shows the potential of automatic image analysis by enabling the processing of the single page in less than a minute. Moreover, the usage of a computer cluster will allow the highly performant processing of large amounts of data. The software framework itself will be integrated as a service into DARIAH infrastructure to make it adaptable for wider range of documents.

14:05-14:30: SemToNotes and eCodicology – Manual and an Automatic Manuscript Annotation Tool Working Together *(Jochen Graf, Cologne)*

Considering the amount of digitized manuscripts available on the web, there is an urgent need to have powerful tools available for the automatic analysis of manuscript pages, particularly too in the light of codicological aspects. Such tools provide us techniques to cluster pages, e.g., with similar layout features, even if the pages are dispersed over large manuscript collections; and perspectively they can perhaps predict the provenance of books by means of feature extraction techniques, even for such books having no metadata available yet at all. Automatic techniques scale well whilst the manual analysis of the growing amount of digitized manuscripts has become unattainable.

It is not clear yet, though, how far codicologists—one of the supposed winners of such automation efforts—can give credence to automatic results. The results seem valuable and indispensable to them, but the epistemic status of the results stays still somewhat unclear just as well. Since few years, there is a widely accepted methodical consensus in the

humanities, which does not assume a complete paradigm shift ending up the qualitative era of analysis and opening up a completely new quantitative age, but rather sees the methodical future of the Digital Humanities in a sensible cooperation between qualitative and quantitative methods, respectively between automatic and manual analysis tools. The linguists are perhaps the pioneers in developing such tools. Besides implementing automatic text mining services, at least to the same extent, efforts are spent in integrating those services into interactive user interfaces to further interpret automatic results, to contest their relevance and validity and to directly make them available for manual correction and extension.

Within the second phase of the project DARIAH-DE (March 201 4 to February 201 6), there is currently the great chance to establish such a semi-automatic service for the codicological domain. The automatic manuscript analysis tool of the eCodicology project is integrated into an interactive manual annotation tool developed at the University of Cologne, called Semantic Topological Notes (SemToNotes). In contrast to similar semi-automatic linguistic tools, the focus here is also on the non-textual data involved, which means, image annotations are not only interpreted as just a link to the actually more important textual information, but the geometry of an image annotation itself—its size, its format or its spatial position on the manuscript page—is of interest as well.

In this non-textual context, it is currently evaluated how standards such as the Text Encoding Initiative (TEI), the Shared Canvas data model or the CIDOC-CRM can serve as viable encoding formats and as interchange formats in a codicological, semi-automatic annotation environment. Finally, the implementation of a topological retrieval system is investigated, which offers the possibility to analyse the semantics of image annotations by means of a combination of coordinate data and ontology-based textual data. A semantic-topological retrieval system is thereby assumed to be the key feature forming a bridge between automatic pattern recognition techniques and concrete human research questions.

*14:45-15:45:* *Coffee Break and Poster Session*

## *Session II: Pattern Recognition in Different Scientific Fields*

*Moderation: Rainer Stotzka*

15:45-16:15: Medical Imaging with 3D Ultrasound Computer Tomography for Early Breast Cancer Diagnosis *(Nicole Ruiter, Karlsruhe; Torsten Hopp, Karlsruhe)*

3D Ultrasound Computer Tomography (USCT) is a promising new medical imaging modality for early breast cancer diagnosis. At KIT we developed and tested the first full 3D USCT system aimed at in-vivo imaging. It is based on approximately 2000 ultrasound transducers surrounding the breast within a water bath. The patient lies prone on a mattress with the breast pendulous. During the data acquisition, 10 Mio ultrasound signals (20 GByte) are recorded in approximately four minutes. After signal processing, these signals are used for reflection, attenuation and sound speed image reconstruction using computationally intensive algorithms on multi GPU clusters. Due to the complementary information of the three images and sub-millimeter resolution, USCT is expected to diagnose breast cancer in an early stage with high sensitivity and specificity. It offers high image quality compared to conventional ultrasound and allows reproducible image acquisition.

Recently we conducted a first in-vivo study with ten patients at the University Hospital of Jena, Germany. For comparison MRI images of the same patients were acquired. Image segmentation based on edge detection and 3D surface fitting was performed to remove the background artifacts from the images. This allows a better visual comparison and enables further image processing. As such an image registration was applied to overcome the considerably different breast positioning in MRI and USCT. The registration of MRI to USCT is based on a biomechanical breast model simulating the buoyancy during USCT image acquisition. For intuitive visualization, registered images of both modalities are fused and presented to radiologists. The feedback of patients indicates a convenient patient comfort. The resulting images are promising: compared to the MRI ground truth, similar tissue structures can be identified. The first in-vivo study was successfully completed and encourages for a second in-vivo study with a considerably larger number of patients, which is currently ongoing.

16:15-16:45: Semi-automated Detection of Ground Monuments in Airborne Laser Scan Data (LIDAR) *(Armin Volkmann, Heidelberg; Karl Hjalte Maack Raun, Heidelberg)*

The project, LiDAR based semi-automatic pattern recognition within an archaeological landscape, is focused on adapting and creating semi-automatic procedures for handling and processing LiDAR data within cultural heritage monument detection and large scale cultural heritage management. Particular emphasis is on the implementation of pattern recognition algorithms for semi-automatic detection and management within 3D, 2.5D, and 2D Airborne Laser Scanning data.
The utilization of Airborne Laser Scanning (ALS) data provides several novel approaches for locating and monitoring cultural heritage, especially in areas of logistical complications, e.g. forest, rough terrain, and remote

areas. In order to cope with the huge amount of generated 3dimensional ALS LiDAR point clouds, systematic and semi-automated procedures needs to be defined in order to control and handle these accumulated amounts of otherwise unrestrained information. In doing so, the effort of this project will be focused on pattern recognition algorithms in order to define quantitative methods of handling and processing 3dimensional LiDAR data and subsequent 2dimensional raster by implementing standardized and state of the art systematic and semi-automated approaches for cultural heritage detection and management. Besides state of the art algorithms for automatic procedures of cultural heritage detection and management, the result of the project will also be focused on key technologies regarding data life cycles in joint collaboration with the Software Methods Group (SWM) and the KIT. The joint collaboration will be focused on data sustainability, and optimizing visualization procedures for many different fields of focus.

## Program for 20th February 2015

### *Session III (a): Layout Analysis of Historical Documents*

*Moderation: Philipp Vanscheidt*

09:30-10:00: Problems of Layout Analysis in Medieval Charters
*(Otfried Krafft, Marburg)*

Among the sources for medieval history charters have an eminent rank. The methods of research on them have been developed since the late 17th century and they were refined during the 19th century. There always was a double focus on charters, as there was a division between their outward features and the text itself and its respective content. The main scope was to find out reliable criteria for the identification of forgeries, which can be frequently found among medieval diplomas. For a judgment on that field outward features are central, at least in true or faked originals. Normally the approach is comparative and regards the handwriting, the seals or the material on which the text was written, and sometimes also graphic symbols like monograms.

However, the layout itself was taken into consideration rather seldom as the general appearance of a page can be seen rather an as addition of a certain number of elements. In spite of this complexity I will try to list up what seems crucial for the layout of these charters, especially focusing on factors connected to the distribution between the parts covered with writing/drafting and the blank parchment (e.g. margins and interlinear spaces). Obviously some scribes working for high-ranking authorities like the Roman pope or the emperor were the first ones to create a geometric

outward appearance of their diplomas. Therefore I will refer to some original examples from them as well as to some (possible) forgeries. A lot of the latter show many aberrations in their layout. Not only for this reason layout analysis, be it automatically processed or not, seems to be a necessary way to enlarge our knowledge not only on medieval charters but also on book-pages, as many elements were designed in an analogous way.

10:00-10:30: Document Structure Analysis: Modelling Unseeable Patterns *(Hervé Déjean, Grenoble)*

Simplifying a bit, layout models described in the Document Layout Analysis literature are rather simple: working at the page level, they describe page elements as a possibly recursive organization (graph, tree) of rectangular boxes. But when reading literature from other communities working on documents (codicology, typography, book design), one can find far richer models used to describe documents.

In this presentation, we would like to show how some layout models found in past and modern text layout practices (ruling, grid) can improve document digitization. We will first present these unseeable layout concepts used for manuscripts and printed books, and then explain how they can be of first importance in today's document digitization. In particular, we will show that the traditional concept of type area is a key notion for modeling document layout. We will illustrate this work with several practical usages and evaluations, from OCR improvement to high-level logical segmentation. These examples will highlight the advantage of developing algorithms operating at Document level (and not at the page level).

*10:30-11:00: Coffee Break and Poster Session*

## *Session III (b): Specialized Analysis of Historical Documents*

*Moderation: Philipp Vanscheidt*

11:00-11:30: The Optical Neume Recognition Project (ONRP) – the Development of Search Tool for Neume Notation in Digital images *(Jennifer Bain, Halifax; Inga Behrendt, Tübingen; Anton Stingl, Freiburg)*

Musicologists who work with the earliest neumes constantly struggle to achieve the kind of familiarity with the musical repertory that the medieval scribes and singers had. Often, scholars work backwards, from versions of a melody notated in more recent, pitch-specific traditions to the earlier, unheightened neume forms, with all the uncertainty this

method presents. There have simply been too many neumes in too many patterns to allow a modern musicologist to recognize them in the way medieval choir masters, already aurally familiar with the repertory, would have done. Neumes originally intended to remind must now instruct, and the many nuances in, and modifications of, hundreds of written symbols over hundreds of chant texts have presented tremendous challenges, to date.

Recent developments in Optical Character Recognition (OCR) software, originally designed to identify and attribute meaning to letters and words, is now being applied to one of the earliest neume traditions, originating in the tenth-century scriptoria in St. Gall, Switzerland. The Optical Neume Recognition Project has developed software to scan and process digital images of the medieval antiphoner held in the Stiftbibliothek at St. Gall with the shelf-mark Cod Sang. 390 / 391, identifying its neumes using image recognition techniques, and encoding that information in an XML schema called the Musical Encoding Initiative (MEI). The result is that neumes can now be located on the digital image, both individually and as part of larger neume patterns. Aided by technology, musicologists are now closer to understanding how neumes were used to represent musical gestures, and recognizing when their graphic patterns coincide with melodic ones. At present, searching for combinations of single basic neumes is possible, using software developed for ONRP. This means that new, quantifiable data about notation can be automatically generated, using the search field of our software interface — a welcome contrast to the days when researchers had to count, by hand, the number of occurrences of a particular neume, a neume combination, or a pattern ("formula"). At the moment, text search is also already fully applicable through the integration of meta-data for the manuscript, available from the Cantus Database. The new capacity to search for neume combinations creates new research opportunities; we can now get a far better sense of the frequency of occurrences of certain neume combinations. When queried, the neumes are highlighted on the digital scans of the page. Once the user has located particular occurrences, musicological questions can help to qualify the results. For example, we might ask, does this particular neume appear at a stressed or an unstressed syllable, and at what part of the sentence and melody section? Is it a particular pattern, part of a transition, or a section from a recitation? Some case studies, highlighting results of certain queries to our automated system, will be presented during the paper.

11:30-12:00: Statistical Models for Word Spotting *(Gernot Fink, Dortmund)*

Despite considerable progress in the field of document analysis and recognition, the automatic reading of handwritten texts is still an

extremely challenging task, especially when it comes to historical manuscripts. When the reliable transcription of documents is no longer feasible, approaches for so-called word spotting come into play. These can be considered as specialized versions of image retrieval techniques, where a query word image is searched for in a collection of documents. The most successful methods build on the idea of using statistical image representations mainly in the form of bag-of-features models. This idea can be combined with statistical sequence models, i.e., Hidden-Markov Models (HMMs), used widely in the field of handwriting recognition. In this presentation I will first give an overview of the development of methods in the area of word spotting. Then I will show how the bagof-words principle known from statistical text modeling can be transferred to the domain of image processing leading to Bag-of-Features (BoF) representations. Afterwards, segmentation-free word spotting techniques will be covered which heavily rely on advanced BoF models. As an extension of these with improved capabilities for representing spatial information, I will then present our recently proposed BoF-HMMs which are among the most successful techniques known in the field of word spotting today. The effectiveness of this hybrid approach combining a statistical image model with a statistical sequence model will be demonstrated with experimental results.

12:00-12:30: MultiSpectral Image Analysis for Writer Identification in Ancient Manuscripts *(Robert Sablatnig, Vienna)*

MultiSpectral Imaging (MSI) is a valuable tool for digitizing manuscripts, since it is a non-invasive investigation technique that is capable of enhancing the contrast of the degraded writings compared to normal white light illumination. This talk first shows how MSIcan be used to enhance ancient and degraded writings, to support the task of writer identification. The manuscripts captured, may contain faded out characters can be partly corrupted by mold and hardly legible. Such writings can be enhanced by applying Fisher Linear Discriminate Analysis (LDA) in order to reduce the dimension of the multispectral scan. Since Fisher LDA is a supervised dimension reduction tool, it is necessary to label a subset of multispectral data. For this purpose, a semi-automated label generation step is conducted, which is based on an automated detection of text lines. Thus, the approach is not only based on spectral information but also on spatial information. Next, this talk shows how writer identification and writer retrieval can be performed using the enhanced images. Writer identification is the task of identifying the writer of a document out of a database of Known writers. In contrast to identification, writer retrieval is the task of finding documents in a database according to the similarity of handwritings. The approach presented uses local features for this task. For each document image the features are calculated and the Fisher Vector is generated using the

vocabulary. The distance of this vector is then used as similarity measurement for the handwriting and is used for writer identification and writer retrieval. Therefore, the main goal of our approach is not to identify the writer but to find all documents written by the same writer as a reference document. The proposed method is evaluated on datasets, the experiments show that the proposed methods perform better than previously presented writer identification approaches.

*12:30-13:30:* *Lunch*

## *Session IV: Opportunities for Interdisciplinary Research*

*Moderation: Rainer Stotzka*

13:30-14:00: Digital Humanities at the Ubiquitous Knowledge
*(Carsten Schnober, Darmstadt)*

The presentation will give an overview over digital humanities projects at the Ubiquitous Knowledge Processing Lab (UKP) at TU Darmstadt. This includes research in the fields of linguistics, philosophy and philology, and history science.

Within the LOEWE programme by the Hessian state, the UKP participated in multiple projects concerning linguistics research. These include machine learning for linguistics analysis, profiling the usage of non-canonical linguistic structures in different languages, and the analysis of linguistic properties of collaboratively generated texts such as Wikipedia.

A recently started project called Natur und Staat aims to support and facilitate research in the field of computational philology. In close co-operation with the TU Darmstadt philosophy department, historical German essays published in the early 20 century are analyzed, particularly regarding the use and description of certain so-called “isms”, words that end on “-ism” such as “nationalism” and “communism”.

With regard to history research, the project “Children and their World” is presented. Apart from the UKP, it involves history researchers and information scientists from different institutions in Germany. The project is designed to serve as a template for similar projects in the future. In an exploratory approach and applying user-centered development paradigms, computational methods are adapted and developed to facilitate historic research on a large corpus of German historical textbooks from the years 1850 until 1918. The latter project, “Children and their World” is presented in depth as it represents multiple challenges that exemplify the challenges of many digital humanities projects. It deals with a particularly interesting period from a historical point of view, and investigates novel

research approaches for specific media types: textbooks and juvenile literature. Established, hermeneutic history research methods are combined with computational technologies, and the project also fosters research in computational linguistics and natural language processing aimed at the specific requirements of the project.

14:00-14:30: Investigating Signals on the Page for Use in Identifying Logical Structures in Texts *(Tuomo Toljamo, London)*

Documentary editors strive to establish reliable texts in the form of editions, which contain presentations of faithful transcriptions and select documentary evidence from writing surfaces and material contexts. Editors who have chosen to embrace the digital medium have faced changes in how editions need to be prepared: often informed by TEI guidelines they work with XML editors in marking up structures and enriching texts with additional information by means of text encoding. Reduced costs of disseminating high quality digital images introduced further changes in quick procession: first, editors found themselves able to grant users access to facsimiles; soon, little reason was seen for an edition not to make available all acquired facsimiles; and now, many are highlighting benefits of tighter integration between transcription texts, facsimiles, and content captured within the images. As indicated by prominent screen estates reserved for facsimiles and reports of shifted user expectations, digital photographic facsimiles seem to have drifted towards the centre, both on-screen and in importance. In part spurred by these developments, scholars have voiced differing views as to how facsimiles should be viewed and what importance one should posit on them. It has been stressed that one needs to have a level of literacy in using them and to understand that, from a pragmatic viewpoint, they are good for some things but poor for others. Beside these discussions, however, it seems undeniable that digital facsimiles readily lend themselves to enabling new possibilities in the form of computational approaches. Acknowledging these developments, this talk describes a curiosity-driven, exploratory research project aiming to investigate the extent to which divisions in logical structures within texts can be automatically identified and encoded by utilising explicit and implicit mark-up of the written page. Partly driven by hypothesised benefits from documentary recontextualisation of facsimiles as informed by the original, features of mark-up can include both visual signals (e.g. topological arrangement and collocation of text on the writing surface, decorated initials in manuscripts), and textual signals of content and form (e.g. key words, punctuation and capitalisation schemes). These features can be seen to be explicit on the page, but they can also be read as implicitly pointing towards particular readings of logical structures over others. The talk presents the research premise and underlying ideas, and focusses on points of departure from most other ongoing work and on potential

disconnects between editors' needs and wants and tool development. The research project is carried out in the Digital Scholarly Editions Initial Training Network (DiXiT) which is funded under Marie Curie Actions within European Commission's 7 Framework Programme (Grant Agreement no. 317436).

*14:30 – 15:00: Closing Discussion*