The European Association for Digital Humanities (EADH)
Association for Computers and the Humanities (ACH)
Canadian Society for Digital Humanities / Société canadienne des humanités numériques (CSDH/SCHN)
centerNet
Australasian Association for Digital Humanities (aaDH)
Japanese Association for Digital Humanites (JADH)

Digital Humanities 2016

Conference Abstracts

Jagiellonian University
&
Pedagogical University

Kraków
11–16 July 2016

Kraków 2016

Edited by

Maciej Eder
Jan Rybicki

DHConvalidator service

Marco Petris

On-line abstracts

Michał Woźniak

Design and typesetting

Maciej Eder

Proof-reading

Aleksandra Ptasznik
Karolina Eder
Daria Wyka
Karolina Wróbel
Weronika Ślęczkowska
Anita Uryga
Sandra Romanowicz

ISBN 978-83-942760-3-4

are considered by allowing for ongoing iterations of analyzing the research material and creating the semantic graph in a formalized and unformalized way (enrichments, annotations, text descriptions). With semantic browsing, aggregations of information, annotation and querying the semantic graph, aspects of close and distant reading are addressed, thus offering new techniques for grasping the research material for qualitative research.

For the Digital Humanities, the focus on mattering of apparatuses offers the possibility to open the design space for digital tools to the diversity of epistemological practices in Humanities. Thereby, an engagement with the diversity of the Humanities comes to the front, enhancing the accountability of boundaries and possibilities of epistemological apparatuses in Digital Humanities.

Acknowledgements

The authors would like to thank the research group around Semantic CorA, especially Marc Rittberger, Lia Veja, Kendra Sticht, Anne Hild, and Anna Stisser. The initial realization of the research environment Semantic CorA was supported by the German Research Foundation (DFG) and its further development is supported in the context of CEDIFOR by the eHumanities program of the German Federal Ministry of Education and Research (BMBF) no. 01UG1416C.

Bibliography

Barad, K. (2007). *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Duke University Press.

Bauer, M. W. and Aarts, B. (2007). Corpus Construction: a Principle for Qualitative Data Collection. In Bauer, M. W. and Gaskell, G. (Eds.), *Qualitative Researching with Text, Image and Sound: A Practical Handbook*. London: Sage, pp. 19–37.

Drucker, J. (2012). Humanistic theory and digital scholarship. *Debates in the Digital Humanities*, pp. 85–95.

Geertz, C. (1973). *The Interpretation of Cultures: Selected Essays*. Basic books.

Grassi, M., Morbidoni, C., Nucci, M., Fonda, S. and Piazza, F. (2013). Pundit: augmenting web contents with semantics. *Literary and Linguistic Computing*: fqt060 doi:10.1093/llc/fqt060.

Love, H. (2013). Close Reading and Thin Description. *Public Culture*, **25**(371): 401–34.

Manovich, L. (2011). Trending: the promises and the challenges of big social data. *Debates in the Digital Humanities*, pp. 460–75.

Moretti, F. (2013). *Distant Reading*. Verso Books (accessed 31 October 2015).

Ramsey, S. and Rockwell, G. (2012). Developing Things: Notes toward an Epistemology of Building in the Digital Humanities. In Gold, M. K. (Ed), *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press, pp. 75–84.

Rockwell, G., Brown, S., Chartrand, J. and Hesemeier, S. (2012). CWRC-Writer: An In-Browser XML Editor. *Poster Presented at Digital Humanities*.

Star, S. L. (1998). Grounded Classification: Grounded Theory and Faceted Classification. *Library Trends*, **47**(2): 218–32.

Venturini, T. and Latour, B. (2010). The social fabric: Digital traces and quali-quantitative methods. *Proceedings of Future En Seine*, pp. 30–15.

Notes

[1] http://thepund.it
[2] http://www.cwrc.ca
[3] http://semantic-cora.org
[4] http://semantic-mediawiki.org/
[5] http://bbf.dipf.de/digital-bbf/spo
[6] Additionally, the technological platform itself offers the possibility to link elements of the graph to established Semantic Web vocabularies (i.e., BIBO or DC); this has been performed for a large part of the data.

Making Sense of Illustrated Handwritten Archives

Lambert Schomaker
l.r.b.schomaker@rug.nl
ALICE, University of Groningen, The Netherlands

Andreas Weber
a.weber@utwente.nl
STePS, University of Twente, The Netherlands

Michiel Thijssen
thijssen@brill.com
BRILL, The Netherlands

Maarten Heerlien
maarten.heerlien@naturalis.nl
Naturalis Biodiversity Center, The Netherlands

Aske Plaat
aske.plaat@gmail.com
LIACS, Leiden University, The Netherlands

Siegfried Nijssen
s.nijssen@liacs.leidenuniv.nl
LIACS, Leiden University, The Netherlands

Fons Verbeek
f.j.verbeek@liacs.leidenuniv.nl
LIACS, Leiden University, The Netherlands

Michael Lew
lewmsk@gmail.com
LIACS, Leiden University, The Netherlands

Eulalia Gasso Miracle
Eulalia.GassoMiracle@naturalis.nl
Naturalis Biodiversity Center, The Netherlands

Katy Wolstencroft
k.j.wolstencroft@liacs.leidenuniv.nl
LIACS, Leiden University, The Netherlands

Ernest Suyver
suyver@brill.com
BRILL, The Netherlands

Bart Verheij
Bart.Verheij@rug.nl
ALICE, University of Groningen, The Netherlands

Marco Wiering
m.a.wiering@rug.nl
ALICE, University of Groningen, The Netherlands

Rene Dekker
Rene.Dekker@naturalis.nl
Naturalis Biodiversity Center, The Netherlands

Joost Kok
joost.n.kok@gmail.com
LIACS, Leiden University, The Netherlands

Lissa Roberts
l.l.roberts@utwente.nl
STePS, University of Twente, The Netherlands

Jaap Van den Herik
jaapvandenherik@gmail.com
LCDS, Leiden University, The Netherlands

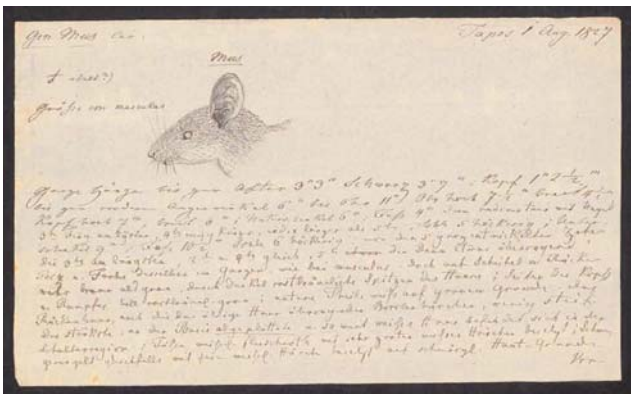

Figure 1. Page from a bundle of field notes, describing and depicting a mouse species. Source: Naturalis Biodiversity Center, Archief van de Natuurkundige Commissie voor Nederlands-Indië. Copyright: Public Domain Mark 1.0

The MONK system uses shape-based feature vector methods that have very few assumptions concerning the content or style of the material. It avoids the traditional OCR approach (optical character recognition) which assumes that individual characters are essentially legible. That assumption only holds for a tiny fraction of handwritten material and a limited number of scripts. The only assumptions MONK makes are that pictorial and textual segments are separated by white spaces; and that the layout, of underlining, etc. in a specific document, is consistent throughout the document. In MONK, classification methods are used that allow for a fast bootstrap from single example instances (nearest-neighbor search) (Gast et al., 2013). With larger numbers of labeled examples, models can be computed, varying from nearest-centroids to support-vector machines and neural networks in a continuous learning process (Krizhevsky et al., 2012; Liu et al, 2015; Guo, in press). A challenging topic from the technical point of view is the relation between existing semantic knowledge (ontologies) and the statistically inferred semantics using Google's *word2vec* and current deep-learning neural networks. Can the underlying structure and style in a collection of a common and realistic size be detected by such algorithms? Can the proposed enrichment system profit from generally available text corpora? The processing power required by the proposed architecture is substantial. For this project, algorithmic optimization of the image processing and recognition system is necessary in order to create the necessary speed and flexibility of the system for use by non-technical end users. In order to tackle this challenge the consortium will make use of the combined knowledge and expertise of ALICE in Groningen, and the Leiden Institute of Advanced Computer Science (LIACS), where multiple supercomputers and high performance computing experts are present.

Because of its visual approach, MONK can handle the diversity of material that we encounter in our use case and in historical collections in general: text, drawings, and images. MONK also does not require a language model nor fully transcribed samples to quickly assess the contents of an archive page. The human-in-the-loop approach of MONK is currently 'label' oriented, but will be enhanced by providing the user and the system with ontologies for bootstrapping the learning process. The system will understand handwritten corpora to such an extent that the visual and textual content on individual pages is categorized, determined and networked to other pages in the archive and external sources. To construct training examples for MONK, biologists and historians of science will manually label documents to the machine learning software by means of a human in the loop approach. In addition, a crowdsourcing approach will be used to further expand this corpus of examples. Our consortium will here build on the expertise of ALICE and Naturalis Biodiversity Center, the Leiden-based National Museum of Natural History. Eventually, the computer-assisted recognition of words and visual information on a page will thus allow users to search, filter and group arbitrary archive items and enables

connections with external databases. Last but not least, MONK lays the groundwork for full transcription of any handwritten-illustrated archival collection.

Figure 2. Drawing of Burro multicolor created in Buitenzorg, Java in 1827 by Pieter van Oort. Source: Naturalis Biodiversity Center, Archief van de Natuurkundige Commissie voor Nederlands-Indië. Copyright: Public Domain Mark 1.0

The central use case of our research project is the collection of the *Natuurkundige Commissie voor Nederlandsch-Indië* (hereinafter NC). It is one of the top-collections of Naturalis. From 1820 to 1850, the NC charted the natural and economic state of the Dutch East Indies and returned a wealth of scientific data and specimens which are now stored in archives in the Netherlands and Indonesia. The collection comprises thousands of handwritten notes and drawings and tens of thousands biological and geological specimens. While these archival items have all been digitized, the individual pages in the notebooks, diaries and reports are not catalogued nor labeled separately. Many of the field notes combine different textual and visual elements on one page. Our short paper presentation is based on the processing of an initial set of understudied handwritten field notes which we carried out in early 2016. By doing so, we will demonstrate the efficiency of the MONK system and our approach.

Owing to the different 'hands' and languages used in the documents, links across handwritten field records and notes, drawings and specimens cannot be made in an efficient way. Our corpus contains material from at least seventeen different writers and the used languages range from German *(Kurrentschrift)* to Latin, French, and Dutch. The labels of related historic specimens only provide very general information on collection localities and collectors. Hence, the typical use case of a scholar wishing to retrieve information on a certain species, person, drawing, or collecting locality is limited. Owing to its sheer dimension and its weak structure, it is impractical to disclose and network this archive manually. Its current inaccessibility hampers research into Southeast Asian natural history and the history of (scientific) knowledge production. Knowledge extracted from the documents will be structured and served as Linked Open Data. This will allow interlinking of content and also enable interoperability with other cultural heritage resources, for example, the physical specimens obtained during expeditions, or other historically significant data collections from the same area.

The multi-layered character of the material makes it a perfect use case for developing a technologically advanced and usability engineered digital environment for interpreting and connecting illustrated-handwritten collections. In our consortium data scientists from the Universities in Leiden and Groningen work closely together historians of science from the University of Twente and taxonomy experts from Naturalis. Fueled by an investment from BRILL publishers in a national funding scheme, this project will not only result in the disclosure of the NC archive, but will also enable the integrated study of underexplored scientific manuscript collections and archives in general.

Bibliography

Zant, T. van der, Schomaker, L. and Haak, K. (2008). Handwritten-Word Spotting Using Biologically Inspired Features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30**(11): 1945–57.

Oosten, J.-P. van and Schomaker, L. (2014a). Separability versus prototypicality in handwritten word-image retrieval. *Pattern Recognition*, **47**(3): 1031–38. (Handwriting Recognition and Other PR Applications)

Oosten, J.-P. van and Schomaker, L. (2014b). A Reevaluation and Benchmark of Hidden Markov Models. *2014 14th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. pp. 531–36.

Gast, E., Oerlemans, A. and Lew, M. S. (2013). Very large scale nearest neighbor search: ideas, strategies and challenges. *International Journal of Multimedia Information Retrieval*, **2**(4): 229–41.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*. pp. 1097–105.

Liu, Y., Guo, Y., Wu, S. and Lew, M. S. (2015). DeepIndex for Accurate and Efficient Image Retrieval. *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. (ICMR '15). New York, NY, USA: ACM, pp. 43–50.

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S. and Lew, M. S. (2015). Deep learning for visual understanding: A review. *Neurocomputing*. (Available online 26 November 2015).