The 3rd HistoInformatics Workshop: Introduction

Marten Düring[1], Adam Jatowt[2], Johannes Preiser-Kapeller[3] and Antal van den Bosch[4]
[1] University of Luxembourg
[2]Kyoto University
[3]Austrian Academy of Sciences
[4]Radboud University Nijmegen
marten.duering@uni.lu, adam@dl.kuis.kyoto-u.ac.jp,
Johannes.Preiser-Kapeller@oeaw.ac.at, a.vandenbosch@let.ru.nl

Abstract

The 3rd HistoInformatics workshop was co-located with the ADHO's DH2016 Digital Humanities conference in Krakow, Poland. 10 out of 17 submissions were accepted. Three recurring areas of interest emerged: Applications and Tools, Historical Research Questions, and Corpus Creation and Management.

1 About the workshop series

HistoInformatics2016, the 3rd International Workshop on Computational History, was held on July 11th 2016 in Krakow, Poland in conjunction with ADHO's 2016 Digital Humanities conference. The HistoInformatics workshop series brings together researchers in the historical disciplines, computer science and associated disciplines, as well as the cultural heritage sector. Historians, like other humanists, show a keen interest in computational approaches to the study and processing of digitized sources (usually text, images, audio). In computer science, experimental tools and methods stand the challenge to be validated regarding their relevance for real-world questions and applications. The HistoInformatics workshop series is designed to bring researchers in both fields together, to discuss best practices as well as possible future collaborations. Traditionally, historical research is based on the hermeneutic investigation of preserved records and artifacts to provide a reliable account of the past and to discuss different hypotheses. Alongside this hermeneutic approach, historians have always extracted data from primary sources and used methods, often borrowed from the social sciences for analysis. A new wealth of digitized historical documents have however opened up completely new challenges for the computer-assisted analysis of large text collections, image corpora and other datasets. Historians can greatly benefit from the advances of computer and information sciences which are dedicated to the processing, organization, and analysis of such data. New

M. Düring, A. Jatowt, J. Preiser-Kapeller, A. van den Bosch (eds.): Proceedings of the 3rd HistoInformatics Workshop, Krakow, Poland, 11 July 2016, published at http://ceur-ws.org

computational techniques can be applied to help verify and validate historical assumptions. We call this approach HistoInformatics, analogous to Bioinformatics and Chemoinformatics which have respectively proposed new research trends in biology and chemistry. The main topics of the workshop are: (1) supporting historical research and analysis, in general, through the application of computer science theories or technologies, (2) analysis and re-use of historical texts, (3) visualisations of historical data, and (4) provision of access to historical knowledge. HistoInformatics workshops took place twice in the past. The first, HistoInformatics2013, was held in conjunction with the 5th International Conference on Social Informatics in Kyoto, Japan in 2013. The second workshop, HistoInformatics2014, took place co-located with the same conference in the following year in Barcelona. As of 2016, the workshop series will alternate between humanities and computer science conferences, with the goal to further strengthen the bridge between the two fields.

2 The 3rd workshop collocated with DH2016

The HistoInformatics workshop series is open to anyone who works on the intersection between computer science and the historical disciplines. As a consequence, the list of topics remains long and varied:

- Natural language processing and text analytics applied to historical documents
- Analysis of longitudinal document collections
- Search and retrieval in document archives and historical collections, associative search
- Causal relationship discovery based on historical resources
- Named entity recognition and disambiguation
- Entity relationship extraction, detecting and resolving historical references in text
- Finding analogical entities over time
- Network Analysis
- Computational linguistics for old texts
- Analysis of language change over time
- Modelling evolution of entities and relationships over time
- Automatic multimedia document dating
- Simulating and recreating the past course of actions, social relations, motivations, figurations
- Handling uncertain and fragmentary text and image data
- Mining Wikipedia for historical data
- OCR and transcription of old texts
- Effective interfaces for searching, browsing or visualizing historical data collections
- Studies on collective memory
- Studying and modeling forgetting and remembering processes
- Estimating credibility of historical findings
- Epistemologies in the Humanities and Computer Science

We have received a total of 17 submissions, of which 7 were accepted as full and 3 as short papers. We have grouped the accepted submissions in three sessions: “Applications and Tools”, “Historical Research Questions” and “Corpus Creation and Management”. Each of these themes represents distinct goals and techniques. In order to address historical research questions with computational methods, historians need to balance the requirement to categorize data and thereby necessarily loose granularity, with the added value of a higher-level perspective on said data and the observation of

emerging patterns. Applications and tools developed with historians as users in mind require relative ease of use, need to be usable for different data sets, scale well and – crucially – be able to help them do something which is of immediate relevance to their work. The accepted submissions address three chronically difficult problems for historians: How to explore very large collections of text, how to observe subtle changes within these collections and how to represent such complex data visually. Corpora pose their own challenges to historians as new types of analysis require restructuring of data and novel search techniques beyond keyword searches.

Dhruv Gupta, Jannik Stroetgen and Klaus Berberich presented *DigitalHistorian: Search & Analytics Using Annotations*. Their system combines state of the art tools in order to detect temporal expressions in text but also uncertainties, and links them to named entities. The system can furthermore highlight time periods which are of particular relevance for a given entity. Carlos Martinez-Ortiz, Tom Kenter, Melvin Wevers, Pim Huijnen, Jaap Verheul and Joris van Eijnatten are also concerned with distant approaches to large collections of digitized texts. In their paper *Design and Implementation of ShiCo: Visualising Shifting Concepts over Time* they present a system which detects changes in word usage. Notoriously subtle and hard to detect, variations in word use, they argue, may point to larger socio-cultural changes. Florian Windhager, Eva Mayr, Guenther Schreder, Michael Smuc, Paolo Federico and Silvia Miksch present their position paper *Reframing Cultural Heritage Collections in a Visualization Framework of Multiple Space-Time Cubes.* Space-Time Cubes, the authors argue, are a powerful means to combine temporal with geographic information, a challenge which is of particular relevance for historians. *HistSearch – Implementation and Evaluation of a Web-based Tool for Automatic Information Extraction from Historical Text* by Eva Pettersson, Jonas Lindström, Benny Jacobsson and Rosemarie Fiebranz discusses the challenges associated with text analysis on historical text. Their system provides normalization of historical text and thereby creates a link to NLP tools developed for modern languages and offers a thorough evaluation of their demonstrator. Andreas Kuczera's paper *Digital Editions beyond XML – Graph-based Digital Editions* introduces graph databases as a means to better store versions of digital editions and to facilitate overlapping annotations, something which is notoriously difficult for XML annotations. Kumiyo Nakakoji, Yasuhiro Yamamoto and Yusuke Kita's paper *Visual Interaction Design for Experiencing and Engaging with a Large Chronological Table of the History of Kyoto* also addresses the problem to represent highly complex data on events and places in an interactive environment. Vilja Hulden's paper *Visualizing labor and business testimony before Congress* relies on metadata of congressional hearings and highlights the presence of business representatives as lobbyists. Patrick Bos, Huub Wijfjes, Maaike Piscaer and Gerrit Voerman study the Dutch social segmentation, also known as pillarization, in their paper *Quantifying "Pillarization": Extracting Political History from Large Databases of Digitized Historical Media Sources.* Sergio Torres Aguilar, Xavier Tannier and Pierre Chastang develop a novel approach for the recognition of named entities in their paper *Named Entity Recognition Applied on a Data Base of Medieval Latin Charters. The Case of Chartae Burgundiae*. Finally, Aline Deicke and Anna Neovesky present their transformation of data stored in a relational database into a graph database and how this opens up opportunities for social network analysis in their paper *A network Analytical Workflow Involving Relational and Graph Databases: Contextualizing Controversies of the Post-Lutheran Reformation.*

3 Programme Committee

We are very grateful to each of our Programme Committee members for their continued willingness to provide reviews of submissions. Each submission received three reviews, typically by

two computer scientists and one humanist to ensure that both dimensions were covered. The members of the 3rd HistoInformatics Programme Committee were:

- Robert Allen (Yonsei University, South Korea)
- Ching-man Au Yeung (Huawei Noah's Ark Lab, Hong Kong)
- Frédéric Clavert (Paris Sorbonne University, France)
- Roger Evans (University of Brighton, United Kingdom)
- Christian Gudehus (University of Bochum, Germany)
- Pim Huijnen (Utrecht University, The Netherlands)
- Catherine Jones (Centre Virtuel de la Connaissance sur l'Europe (CVCE), Luxemburg)
- Nattiya Kanhabua (Aalborg University, Denmark)
- Tom Kenter (University of Amsterdam, The Netherlands)
- Mike Kestemont (University of Antwerp, Belgium)
- Adam Kosto (Columbia University, USA)
- Günter Mühlberger (University of Innsbruck, Austria)
- Andrea Nanetti (Nanyang Technological University, Singapore)
- Daan Odijk (University of Amsterdam, The Netherlands)
- Marc Spaniol (Max Planck Institute for Informatics, Germany)
- Nina Tahmasebi (University of Gothenburg, Sweden)
- Lars Wieneke (Centre Virtuel de la Connaissance sur l'Europe (CVCE), Luxemburg)

4 Acknowledgments

We would like to thank the organizers of the DH2016 conference for agreeing to host our workshop and for their support, and also the anonymous reviewers who evaluated our proposal, for their support and constructive criticism.