Digitala lärobjekt för det omvända klassrummet inom humaniora

Hietanen, Mika

2016

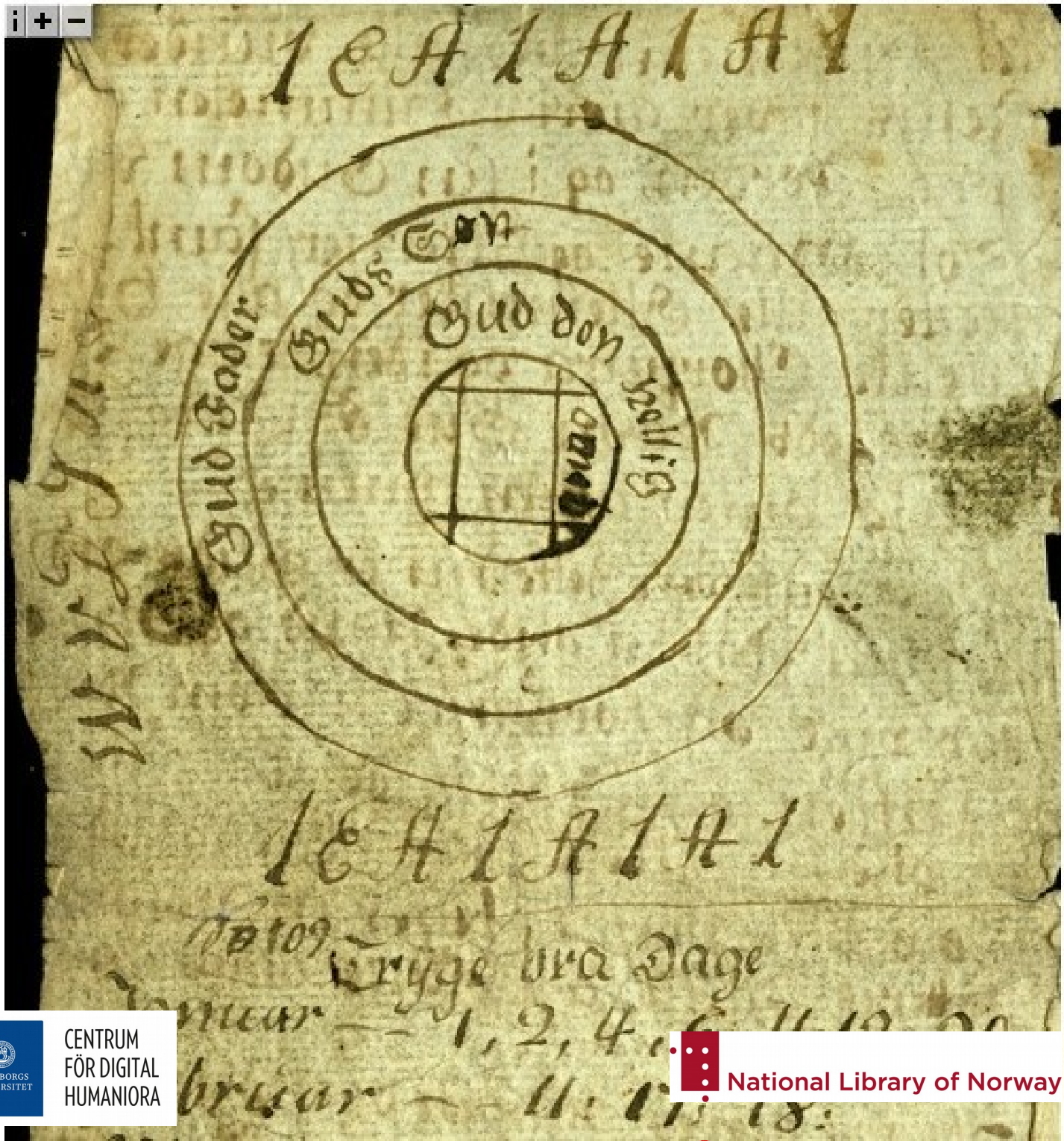
Document Version:
Förlagets slutgiltiga version

DIGITAL HUMANIORA I NORDEN

DIGITAL HUMANITIES IN THE NORDIC COUNTRIES

DHN 2016

Oslo March 15-17



CENTRUM FÖR DIGITAL HUMANIORA

GÖTEBORGS UNIVERSITET

National Library of Norway

UiO : University of Oslo Library
Humanities and Social Sciences Library

UiO : Faculty of Humanities
University of Oslo

DHN 2016

Digital Humanities in the Nordic Countries

Oslo March 15-17 2016

Conference Abstracts

Published by:
UiO, Institutt for lingvistiske og nordiske studier - 2016

ISBN 978-82-90954-51-7

Editor:
Espen S. Ore

Cover: A page from a Norwegian grimoir in the Norwegian Folklore Archives, *NFS Moltke Moe 106 VI b*

International Programme Committee

Chair: Christian-Emil S. Ore, (University of Oslo, Norway)

Haraldur Bernharðsson (The Árni Magnússon Institute for Icelandic Studies, Iceland)

Lone Dirckinck-Holmfeld (Aalborg University, Denmark)

Øyvind Eide (University of Cologne, Germany)

Sakari Katajamäki (Finnish Literature Society – SKS, Finland)

Ditte Laursen (National Library Aarhus, Denmark)

Bente Maegaard (Copenhagen University, Denmark)

Mats Malm (Gothenburg University, Sweden)

Annika Rockenberger (University of Oslo, Norway)

Anne Birgitte Rønning (University of Oslo, Norway)

Local Organizing Committee

Chair: Anne Birgitte Rønning (University of Oslo, Norway)

Federico Aurora (University of Oslo, Norway)

Jenny Bergenmar (Gothenburg University, Sweden)

Line Esborg (University of Oslo, Norway)

Trond Haugen (National Library, Norway)

Lars Jørgen Myking (University of Oslo, Norway)

Espen S. Ore (University of Oslo, Norway)

Hedda S. Rui (University of Oslo, Norway)

Preface

As a scholarly discipline, Digital Humanities has a very short history, compared to other fields of the humanities with their history of several thousand years. On the other hand, we should observe that computers have been used in the humanities since the early 1950ies, that is nearly as long as electronic computers have existed. In 1949 Roberto Busa started his collaboration with IBM on producing a complete concordance for the works of Thomas Aquinas. The project has many of the characteristics of later large projects. It took 30 years to complete, and Busa had to think outside the box and designed job training schemes for young Italians to get the manpower needed. Busa's project is in the field later called literary and linguistic computing. The former name of the European Association for Digital Humanities (EADH) was the Association for Literary and Linguistic Computing (ALLC). Our new Nordic association, organizing this conference, Digital Humaniora i Norden (DHN), is an Associated Organization of EADH.

The connection to the well-established discipline Literary and Linguistic computing is reflected in the programme. There are sessions on literary studies, corpus linguistics, paleography, text philology and digital classics. However, the programme shows us that Digital Humanities is broader than LLC. There are sessions on art history and musicology; history, nations and citizenship; social media and community participation, and on basics such as modelling, simulation, infrastructure and large scale digitization.

There is also a poster session covering a wide range of fields.

The response to the call for paper to DHN2016 was overwhelming. We received 125 proposals for presentation and posters. The final programme consists of 79 paper presentations and 13 posters, all presented in this book of abstracts. Digital Humanities in the Nordic countries is indeed an active flourishing activity. We wish to give our warmest thanks to the Programme Committee and the reviewers who helped us select the presentations included here.

Christian-Emil Ore, Chair of the Programme Committee
Anne Birgitte Rønning, Chair of the local Organizing Committee

Table of Contents

Preface....7
Semantic Web as an interlinking environment for knowledge dissemination in the cultural heritage domain....15
Can there be a Nordic digital humanities?....15
Documenting folklife in the digital age....16
DIGITALE KULTURHISTORISKE ARKIV – hvorfor og hvordan?....19
The Institute for Language and Folklore....19
Historical Accounts, Commodities, and Services: Descriptions and Methods for Digital Humanities Questions....21
Modeling semantically Enhanced Digital Edition of Accounts: A Report on the MEDEA Project...22
A Wealth of Choices: Situating Economic Information in Digital Scholarship....23
A Review: Heinrich Libler's Report from the year 1640: Ideas on connecting historical economic Data to modern digital Accounting....24
Massedigitalisering og digital humaniora....27
Temamodellering....27
«The Rebirth of the Author»: Eller hvordan å gjenkjenne litterær stil ved hjelp av dataanalyse....29
Dagny Juel Przybyzsewska i *Samtiden*. Fjernlesning som kontekstualiseringsstrategi....30
Ett komplett arbetsflöde för två digitala utgåvor....31
Arbetsflödet och de digitala verktygen inom två textkritiska utgivningsprojekt....31
Användning och anpassning av verktyg inom ZTS och HPS....33
Anna Movall/Elisa Veit....33
Ett brev annoteras – en demonstration av arbetsprocessen....36
Eliel Kilpelä....36
Spatial Humanities....41
Spatial Turn, Texts and Mapping – GIS in the Humanities....42
From Iceland to Jerusalem - the Itinerary of Nikulás of Munkaþverá as a web edition....43
The Norwegian Folklore Archive and narratives of mapping....44
Södra Råda: a virtual diorama in time-space....47
Narrative Approaches to the Digitalization of Participatory Urban Planning:....48
Organizing complexity. On some challenges in the linguistic annotation of Mycenaean Greek texts. 49
Reception history across languages – a challenge for the digital humanities....51
Tidens Stroem – the flow of time....52
Finländska klassikerbiblioteket – finländsk litteratur i det digitala landskapet....54
PRINSIPALKOMPONENTER - PRINCIPAL COMPONENTS....55
A nation's web over time: The Development of the Danish Web 2005-2015....56
Nordic Englishes on Twitter....57
The blindspot of Digital Humanities or What Ethnography can contribute....58
How (not) to do textual criticism online: A user's perspective on selected Nordic digital scholarly editions....60
Digital Maps of Scandinavian Literature in Italy (Lt.it): a Bibliographic database of Italian translations of Scandinavian literature....61
Transcription by dictation....62
Geographical mapping of *Sagnagrunnur*, a database of published Icelandic legends....63
Modeller av mennesker, mennesket som modell....65
Streaming Cultural Heritage: Following Files in Digital Music Distribution....67
Historians digging in the text mine....67

Integrating heterogeneous digital text resources to analyse complex neologisms in Norwegian.....68
Counterfactual history in simulation video games: a methodological approach to studying historical consciousness.....70
Using electronic maps to show speech data.....72
Perceptions of Truth/Untruth in Ibsen Criticism.....73
Digitala lärobjekt för det omvända klassrummet inom humaniora.....73
Ibsen and the Digital Humanities.....75
The Creation and Features of the Online Archive Devoted to Sigurður Guðmundsson, the Painter, and the "Evening Society".....76
The Musical Gestures Toolbox – Towards Digital Musicology in Matlab.....77
Dealings with Uneven Corpus – Experiences from the Use of a Difficult Research Data.....78
The Many Shapes of Vladimir Nabokov's 'Lolita' and 'Mademoiselle O': a Quantitative Analysis 80
Exploring Scholarly Craftsmanship and Digital Research Methods in nodegoat.....81
Assessing lexical quality of a digitized historical Finnish newspaper collection with modern language technology tools.....83
Unlocking a Finnish Social Media - In Search of Citizen Mindscapes.....85
Icelandic Saga Map Project: Digitally Mapping Medieval Icelandic Literature.....86
Mapping life-stories of the Moravian Church 1730-2000.....87
Digital Palaeography and the Old Swedish Script. Some Preliminary Results and Future Prospects 89
Emergent Environments.....90
Cross border access for researchers to source material by ECL.....92
Uses and Extensions of Digital Resources for Multimodal Communication in an International Study Program at the University of Copenhagen.....93
A cross cultural examination of the correlation between 'selfies', self-esteem and introversion in young people living in Ireland and Hong Kong.....96
Making web pages for TEI-P5 texts and connecting them with dictionaries.....101
Hur kan digitala samarbeten mellan forskare och minnesinstitutioner underlättas?.....102
The Danish Thesaurus: semantic and thematic structure and its use in digital humanities.....103
Mapping Prehistory.....106
Kan digitale utgaver være vitenskapelige utgaver?.....108
The Copenhagen Associations Project.....110
Coping with disaster: the impact of Stórabóla on abandonment, reoccupation and land use in early 18th century Iceland.....111
Travelling TexTs: A HERA financed five countries project from the point of view of its Finnish team 111
The ELMCIP Project, Electronic Literature Knowledge Base, and Data Visualization Research Outcomes.....112
De svenska medeltidsbreven som digital resurs.....114
Data-driven text analysis for digital humanities: some thoughts on how and why.....115
Swedish Women Online (SWO) from the Middle Ages to the Present.....117
Codicologist vs. Computer: How can Machine Learning Support Research in Manuscript Studies? 119
Project presentation: The Jihadi Document Repository.....120
Locus Classicus.....121
SWE-CLARIN – the Swedish CLARIN project – aims and activities.....122
Manart – en databas som gör konsthistoriens dynamik *manifest*.....124
Teaching Digital Humanities at the Faculty of Arts at the University of Helsinki.....126
Conjuring up the Artist from the Archives: Ivar Arosenius.....127
One for all - How to transcribe Icelandic manuscripts for scientific research.....129

Länkade öppna data och den antika världen – två svenska forskningsprojekts deltagande i internationella samarbeten....130
Digital Editing with Virtual Infrastructures: TextGrid and DARIAH-DE....135
Kulturminnena talar – med hjälp av språkteknologiska metoder....136
Classical Intertextuality in Late Greek Poetry: a Computational Approach....137
Comparative palaeography of cuneiform: a new digital humanities method and approach....139
Att hantera felegrafen: Textanalys av smutsiga tidningar....139
Digital Humanities in the field of Linguistics and Language Research....141
From Handwritten Slips to a Digital Dictionary Corpus....142
Hassle-free creation of metadata and conversion of text material to TEI format....144
Scribes and Late Latin Charters: a treebank-based study of the language competence of scribes...145
Setting up a Digital Humanities Centre in Romania. Problems and solutions....147
GROUP INTERACTION AND SYNCHRONISATION TO MUSICAL PASSAGES OF ELECTRONIC DANCE MUSIC....147
Digitizing And Opening Access To Six Icelandic Dictionaries....149
Digital Humanities in Swedish research: a bibliographic review....150
Henrik Ibsen - in text and on stage....151
Utvikling av konseptet “LÆR KIDSA MUSIKKKODING”....154
SEMATIA: Linguistic Annotation of Greek Papyri....155
Repertory of Conjectures to Horace....156
Digital Classics. A hands-on Introduction to EpiDoc and Treebanking....159
Bukvik - a DH Scholar’s Environment for Stylistic Analysis....162
CollaboFramework Framework and Methodologies for Collaborative Research in Digital Humanities....167

Plenary keynote lectures

Semantic Web as an interlinking environment for knowledge dissemination in the cultural heritage domain

Francesca Tomasi
University of Bologna, Italy
francesca.tomasi@unibo.it

Semantic Web technologies have had a pervasive control on research activities in the cultural heritage domain. Each process related to data modeling moves toward the knowledge graph principle. Everything is connected, and the connection defines the real meaning of the observed domain.

Starting from these considerations the aim of this speech is to present some activities in enhancing cultural heritage information starting from the semantic power of data. In particular the process of ontology creation is the key to reflect on building relationships between pieces of knowledge. The discussion starts from the work on Vespasiano da Bisticci Letters as a case of semantic digital edition, to keep on with the Zeri Photo Archive, as an open linked data collection in the art of history domain, to continue until the abstract semantic modeling of ideas - i.e. from the interpretation act in authorship attribution to the role of people, mined from authority records, in an event-centric and context-oriented perspective.

All these activities will be described in the context of the Italian research environment in which I work: our Italian Association of Digital Humanities (AIUCD, Associazione per l'Informatica Umanistica e la Cultura Digitale), the Bologna Digital Libray (AlmaDL) and the centre for curation of digital resources that we built in the University of Bologna (CRR-MM, Multimedia Research Resource Centre).

Can there be a Nordic digital humanities?

Patrik Svensson
Umeå University
patrik.svensson@umu.se

In this talk, I will address three clusters of questions:
1. Is there a Nordic flavor of digital humanities? What does it even mean to imagine a Nordic digital humanities in the context of the organized digital humanities spreading across the world? What is the global perspective and what is local about the digital humanities?
2. What could the digital humanities become through the Nordic digital humanities? How do we make the most of this opportunity? What may be possible directions? What is the intellectual gist? How do we make a difference? What can we learn from others and from our own past?
3. What is the place of the digital humanities? How can we relate the digital humanities to the development of the humanities and the academy in the Nordic countries and internationally? How do we relate to areas such as environmental humanities, gender studies and urban studies?
I will develop an argument based on these questions and suggest a possible direction forward

without prescribing any solutions or templates. In developing that argument, partly building on my new book (*Big Digital Humanities*), I will attempt to outline an intellectual-material basis for the field and discuss what it takes to make the digital humanities at this point in time. I argue that we need to see the digital humanities as a liminal and intersectional endeavor with a commitment to multiple genealogies, epistemic traditions and modes of engagement. DH can be a curated place with sharpness and generosity empowering intellectual pursuit and the development of humanistic infrastructure and arguments.
The talk will engage with a series of case studies, experiences and expressions.

Documenting folklife in the digital age

Nicole Saylor
Library of Congress
nsay@loc.gov

The proliferation of smart phones, tablets and other networked technologies has revolutionized how the cultural record is written. For the American Folklife Center at the Library of Congress, which is mandated to preserve and present folk culture, the implications are profound. A new mobile app has taken oral history collecting efforts global, yielding nearly 100,000 submissions to the archives in the app's inaugural year. As folk expression proliferates across the web, the Library is harvesting a selected set of exemplar vernacular websites. Meanwhile, the task of making analog multi-format ethnographic collections accessible at digital scale creates significant challenges, including the need for tools that enable Indigenous communities to manage their intellectual property and cultural heritage. In this talk, the audience will be asked to work through such curatorial issues that impact digital humanities inquiry. They will be asked to consider ways in which collaboration between digital humanities scholars and cultural heritage workers can lead to mutually beneficial innovation and knowledge creation.

Sessions

DIGITALE KULTURHISTORISKE ARKIV – hvorfor og hvordan?

Line Esborg
University of Oslo
line.esborg@ikos.uio.no

Audun Kjus
Ida Tolgensbakk
Norsk Folkemuseum
audun.kjus@norskfolkemuseum.no
ida.tolgensbakk@norskfolkemuseum.no

Fredrik Skott
The Institute for Language and Folklore
fredrik.skott@sprakochfolkminnen.se

Arrangører:

Norsk Folkeminnesamling, UiO

NFS har siden opprettelsen i 1914 jobbet med å samle, bevare og formidle tradisjonsmateriale. Det dreier seg i hovedsak om UNESCO-listede skriftlige nedtegnelser av muntlig folkediktning som eventyr, sagn og ballader samlet inn på 1800- og 1900-tallet. I tillegg kommer nyere materiale etter for eksempel de store norske livshistorieinnsamlingene i 1964, 1981 og 1996.

Norsk etnologisk gransking, Norsk Folkemuseum

NEG er et tradisjonsarkiv ved Norsk Folkemuseum der det er samlet mer enn 40 000 personlige beretninger om livets ulike sider. Fra NEG ble grunnlagt i 1946 har arkivet stilt de frivillige meddelerne et stort antall spørsmål om emner som hører dagliglivet til.

Dialekt-, ortnamns- och folkminnesarkivet i Göteborg

DAG omfatter et svært variert materiale: ulike former for skriftlig materiale, fotografier, intervjuer, musikkinnspillinger. Det består av folkeminner opptegnet i det vestlige Sverige fra 1800-tallet til i dag, en dialektordsamling og en stedsnavnssamling.

Sammendrag:

Det er lange tradisjoner for kulturhistoriske arkiv i Norge og Norden for øvrig, og feltet er mangslungent og desentralisert. Vi har folkemusikkarkiv, folkloristiske og etnologiske arkiv, dialekt- og stedsnavnsarkiv, små arkiv basert på ulike typer minneinnsamlinger, og så videre. De kulturhistoriske arkivene er oftest levende arkiv. De inneholder omfattende eldre materiale, men driver samtidig nyinnsamling. Både de eldre og de yngre delene av arkivene består oftest av initierte kildetyper som feltarbeid, intervjuer og spørrelister. De kulturhistoriske arkivene er i dag viktige kunnskapsbaser, og er ofte svært vitale. Samtidig er de nesten uten unntak små institusjoner med få ansatte, og i den institusjonsstrukturen de er en del av er de ofte uten støtteapparat tilpasset egne behov.

Foreløpig har de kulturhistoriske arkivene falt mellom stolene i arkiv-bibliotek-museumssektoren. Vårt materiale gjør seg ikke nødvendigvis godt i for eksempel et system som det norske Digitalt Museum. Flere av de nordiske kulturhistoriske arkivene er i dag lite kjent, og viktig eldre materiale lider under vanskjøtsel og glemsel. Vår store fordel i dag er at både vårt materiale og våre arbeidsmetoder kan se ut til å passe som hånd i hanske med digitalisering og sosiale medier. Dette må utnyttes! Det er behov for en gjennomtenkt løsning for både digital lagring, innhenting av materiale og publisering. Vi trenger ikke bare et digitalt arkiv, men en digital arbeidsmåte. Hvorfor og hvordan gjør vi det?

Innledninger til panelet:

Hvis vi får konferansekomiteens velsignelse, ønsker vi å legge vekt mer på debatt enn på foredrag i dette panelet. Det vil si at vi lar presentasjonene vare i 15 minutter, med 15 minutter til spørsmål og diskusjon. Vi planlegger følgende tre korte foredrag:

- ET LEVENDE ARKIV: HVA ER DET? Hvilke unike behov har vi i kulturhistoriske arkivinstitusjoner – som levende, uavslutta arkiv?

  Ved Audun Kjus, NEG

- HVEM EIER ARKIVET? Hva innebærer det om vi ønsker (eller er lovpålagt) å gi deltakere i dokumentasjonsprosjekter full råderett over egne bidrag?

  Ved Ida Tolgensbakk, NEG

- ÅPNE OPP DE GAMLE SAMLINGENE – UPROBLEMATISK? Digitalisering gir fantastiske muligheter for å aktualisere eldre materiale: gjøre det tilgjengelig, og koble det på andre prosjekter. Men er alt like aktuelt? Og hvordan tar vi vare på alle sider av kildeverdien til materialet?

  Ved Line Esborg, NFS

Historical Accounts, Commodities, and Services: Descriptions and Methods for Digital Humanities Questions

Andrew Wareham
Roehampton University
a.wareham@roehampton.ac.uk

Kathryn Tomasek
Wheaton College, Massachusetts
tomasek_kathryn@wheatoncollege.edu

Paige Morgan
University of Miami
p.morgan@miami.edu

Monika Eisenhauer
monikaeisenhauer@freenet.de

Session Abstract

Thematic Question: Given the abundance of economic information that exists in archives and printed texts, how can we optimize capture, sharing, and accessibility for the diverse questions that digital humanities practitioners and other scholars might wish to pose?

During the second half of the twentieth century, economic historians created data sets from some economic sources, but most scholars tended to consider the kind of detailed local economic information found in account books to be too labor-intensive to use extensively. Thus such sources were largely left to social science historians and their sampling methods. In 1998, literary and cultural scholar Mary Poovey brought accounting sources to the attention of Anglophone humanists with her monograph, *A History of the Modern Fact: Problems of Knowledge in the Sciences of Wealth and Society*. And in her subsequent book, *Genres of the Credit Economy: Mediating Value in Eighteenth- and Nineteenth-Century Britain* (2008), Poovey posited that a series of cultural decisions had occurred in Great Britain that led to economic information's becoming valued more highly when it was presented as facts grounded in mathematical calculations than when it appeared in literary form. The advent of the Internet and the WorldWide Web brought opportunities for scholars to present digital editions of accounts of various sorts, and the papers included in this session present three perspectives on opportunities for presenting information about accounts, commodities, and services on the Semantic Web.

The session is linked by the participants' involvement in the MEDEA project (Modeling semantically Enhanced Digital Editions of Accounts), which is currently supported by a Bilateral Digital Humanities award from the National Endowment for the Humanities (U.S.) and the German

Research Foundation (Germany).[*] The papers might be seen to represent a developing process: the first outlines the general questions posed by the MEDEA project; the second describes a present project; and the third introduces a possible future method.

The session's chair presented at the first MEDEA workshop at the University of Regensburg in October 2015.

The author of the first paper is one of the Principal Investigators of the MEDEA project. The paper will describe the objectives of the project and outline the MEDEA team's current recommendations for best practices in creating digital editions of accounts optimized for comparison across time and space.

The author of the second paper discusses an in-process project, an archive of prices for goods, services, and experiences found in printed texts. This project restores price to the reading experience by providing information that was once available to readers but has been lost over time; its aim is to make price information a vehicle for discovery. The author plans to attend the second MEDEA workshop at Wheaton College in April 2016.

The author of the third paper will present recommendations for standardizing a concrete procedure for transforming early modern accounts into annual accounts comparable to twenty-first-century data. The author seeks to develop ideas for rationalization and standardization through the processes of identifying and transforming numerical data found in old account books into useful tools for editors who are creating digital editions. The author presented at the Regensburg workshop.

Papers

Modeling semantically Enhanced Digital Edition of Accounts: A Report on the MEDEA Project

Kathryn Tomasek

A collaborative project that brings together European, Asian, and North American scholars interested in ways the Semantic Web offers opportunities to collect and compare data from multiple digital projects, the MEDEA project looks to the potential of developing broad standards for producing semantically enriched digital editions of accounts. Our project is grounded in the widespread use of the Guidelines of the Text Encoding Initiative (TEI) for producing stable humanities-oriented data from textual sources, and MEDEA explores models for building on the Guidelines to test ways to publish data on commodities, wages, and prices susceptible to comparative analysis.

. This project is supported jointly by the National Endowment for the Humanities and the Deutsche Forschungsgemeinschaft.
The Deutsche Forschungsgemeinschaft (German Research Foundation) is the central, self-governing research funding organization that promotes research at universities and other publicly financed research institutions in Germany. The DFG serves all branches of science and the humanities by funding research projects and facilitating cooperation among researchers.
The National Endowment for the Humanities is an independent grant-making agency of the United States government dedicated to supporting research, education, preservation, and public programs in the humanities.
Any views, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect those of the National Endowment for the Humanities or the Deutsche Forschungsgemeinschaft.

While the TEI Guidelines provide a standard for markup of manuscript and print sources, some of the elements and attributes most useful for accounts fail to model machine-readable values adequate to the goal of comparability of accounts originally created across broad ranges of time and space. Thus the MEDEA project looks also to CIDOC-CRM and RDF/OWL as sites to begin consideration of the kinds of taxonomies and ontologies that will produce standard machine-readable values to express some of the semantic values found in accounts--especially information about commodities and currencies--that are relevant to humanities scholars. This includes conversion of local measures.

The MEDEA project thus looks for ways to provide interconnecting commodities information and aims to enhance the usability of local accounting sources. The anticipated long-range products will enable international researchers and the interested public to share data about accounts originally created across broad ranges of time and space.

The project combines the work that I have been doing on a local set of financial records kept by members of the family that founded Wheaton Female Seminary with work being done under the Chair of Social and Economic History at the University of Regensburg and with the brilliant work of European Digital Humanities scholar Georg Vogeler, who works at the Center for Information Modeling of the Austrian Centre for Digital Humanities in Graz.

At present, the MEDEA leadership team imagines semantically enriched digital editions as networks of references between several digital representations of original archival account books. Such editions will allow scholars with varied interests to use the data from the accounts in different ways according to their fields and scholarly interests. Economic historians, for example, will be able to extract numerical data, business historians commodities information, book historians information about materials stocked by printers and publishers, social historians examples that demonstrate the role of economic exchanges in a community and beyond its borders. All of these scholars will be able to record source references that can be traced to other scholarly representations of the account books—those made by linguists, for example.

To produce such discovery-oriented semantically enriched digital editions of accounts, we consider RDF as a good solution to publish the data extracted from accounts in machine-readable format as it facilitates explicit references to other (possible) representations of the accounts. The paper will present a brief overview of recommended tools and formats that will be tested in preparation for a second MEDEA workshop in April 2016.

A Wealth of Choices: Situating Economic Information in Digital Scholarship

Paige Morgan

Visible Prices (VP, http://www.visibleprices.org) is an in-process project to develop an archive of prices for goods, services, and experiences, collected from literary & historical sources, currently focusing on eighteenth & nineteenth-century England, with the potential to cover larger chronological and geographical ranges. For example, in Charlotte Bronte's Jane Eyre, Jane's salary as governess is £30 per annum. VP makes price back into a visible part of the reading experience by allowing users to discover what £30 meant in purchasing power, or to see how the salary that Bronte chose compared with those being offered in nonfictional sources. VP allows researchers to

see prices from a different perspective than mathematical inflation calculators, to better understand how prices were used as information, and to learn how the evolving use of information shaped economic and cultural development. It captures individual prices against a panoramic backdrop, allowing users to glimpse economic information in broader contexts. Thus, it helps to open the study of money and commerce to disciplines outside of economics. VP is a response to the influx of historical data from printed sources onto the web. Aggregating that wealth of data presents disciplinary challenges – how to be both rigorous and permissive in creating a resource that might be used by humanities and social science researchers; and technical challenges – how to effectively and accurately gather big data, and choose an encoding system for a highly variant corpus of texts.

In this talk, I will focus on the work of creating data and metadata for entries in Visible Prices. Because price information appears in such a wide variety of texts (from printed advertisements to municipal documents to personal correspondence), any single price will be accompanied by different densities of information. One price may include details about the precise geographical location where it was offered, the quantity being sold, and where goods being sold originated, while another price may simply be recorded as "Almonds, 6 shillings." My goal with Visible Prices is to stay focused on the prices -- and by doing so, make them a vehicle for discovery. However, the decisions I make when structuring information for the database are significant because they connect price information with other fields, for example gender and class studies, or studies around a particular industry (i.e. textiles, agriculture, etc.). These decisions impact the durability of the project -- too much metadata makes for a fragile and unwieldy database -- and also, how accessible it is to outside researchers -- too little metadata results in a database where the answers to research questions are always just out of reach. This leads to a key question: what is the most important information to include when digitizing prices? As I answer this question in regards to Visible Prices, I will also address how we might identify essential information for other digital humanities database projects built around specialized subjects.

A Review: Heinrich Libler’s Report from the year 1640: Ideas on connecting historical economic Data to modern digital Accounting

Monika Eisenhauer

The report of Heinrich Libler from the year 1640 describing economic data of the Early Modern Period can be found in the legislative records of the Bursfelde Congregation. This monastic congregation was founded in 1446 as a union of the reformed Benedictine monasteries in the German regions. For many centuries the number of its members had been growing according to the rate of reforms occurring in the abbeys with the intention of renewing monastic life. Simultaneously the organization’s structure strengthened and formed a compact body over the years. The consolidation of the administration structure can be read off the recorded economic data.

Libler’s report belongs to the recorded documentation of the annual general assembly and presents several years of economic activities of the Bursfelde Congregation. As presented in Regensburg the revenues and expenses of the economic years 1628-1640 can be extracted from the source and transformed into a modern cameralistic account by using 1. identification, 2. transformation and 3. analysis. Exercising the tools of relativity and extrapolation we get an annual account, which is comparable to modern day’s data.

But how can we standardize this concrete procedure to get an automated operation? How can we embed this process into given digital tools?
To develop ideas about the process it is beneficial to focus on the identification and the transformation of data – the analysis after the transformation is already described by modern economics.

1. ***Identification***: After identifying the data as *economic* data it is necessary to make broader differentiations: The data has to be categorized by relating them either to the operational accounting or to the financial accounting. Within the context of digitization this leads to the request of automating this process: Can we create something like a "digital assistant" that helps us quickly to gain an equivalent in modern accountancy?
2. ***Transformation***: A high percentage of historical economic data refers to the financial accounting. As exposed in Regensburg the best frame of reference may be the modern cameralistic accountancy, because its simplicity guarantees the most effective equivalence. But according to the purpose of digitalizing we surely do not aim at having a manual data extracting. That leads to the following considerations:
   1. How can we connect programs like TEI or CSM to programs/data banks of accountancy? Which partition of the definition 'measurement' could help to transfer information from one data bank into another?
   2. Which categories of commodities could be useful given the aim of connecting data of the Medieval Ages and the Early Modern Times to modern book-keeping systems?

My talk aims at developing ideas of rationalization and standardization in the process of identifying and transforming the numeric data found in old account books into useful tools an editor can handle while creating a digital edition.

References:

CIDOC-CRM <http://www.cidoc-crm.org/>

Nick Crofts, Martin Doerr, Tony Gill, Stephen Stead, Matthew Stiff (editors), **Definition of the CIDOC Conceptual Reference Model,** December 2011. ISO 21127:2014

Debt- and Invoice-Registers of the Teutonic Order around 1400 <http://www.schuredo.uni-hamburg.de/content/below/index.xml>

MEDEA (Modeling semantically Enriched Digital Edition of Accounts) <http://medea.hypotheses.org/>

Poovey, Mary. A history of the modern fact: problems of knowledge in the sciences of wealth and society. Chicago: University of Chicago Press, 1998.

Poovey, Mary. Genres of the credit economy: mediating value in eighteenth- and nineteenth-century Britain. Chicago: University of Chicago Press, 2008.

RDF Working Group, Resource Description Framework (RDF), < http://www.w3.org/RDF/ > .

Robert Sanderson and Benjamin Albritton, Shared Canvas Data Model
< http://iiif.io/model/sharedcanvas/1.0/index.html > .

Jürgen Sarnowsky (lead researcher), *Die mittelalterlichen Schuldund Rechnungsbücher des Deutschen Ordens um 1400* < http://www.schuredo.unihamburg.de/content/ > .

TEI Consortium, eds. *Guidelines for Electronic Text Encoding and Interchange* . [Last updated on 15th October 2015]. < http://www.teic.org/P5/ >.

Kathryn Tomasek and Syd Bauman, «Encoding Financial Records for Historical Research», *Journal of the Text Encoding Initiative* 6 | December 2013; DOI : 10.4000/jtei.895 [online] .

Visible Prices <http://www.visibleprices.org>.

Georg Vogeler: «Warum werden mittelalterliche und frühneuzeitliche Rechnungsbücher eigentlich nicht digital ediert?», *Zeitschrift für digitale Geisteswissenschaften* 1 | January 2015; DOI: 10.17175/sb001_007 [online]

W3C, OWL Web Ontology Language
< http://www.w3.org/standards/techs/owl#w3c_all >.

W3C, OpenAnnotation Data Model < http://www.openannotation.org/spec/core > .

XBRL Global Ledger: Transactional Reporting <https://www.xbrl.org/the-standard/what/global-ledger/>

Massedigitalisering og digital humaniora

Lars G. Johnsen, Trond Haugen, Arne Martinus Lindstad, Magnus Breder Birkenes
Nasjonalbiblioteket

Anne Birgitte Rønning
University of Oslo

Store digitaliseringsprosjekter, kombinert med de riktige verktøyene for behandling av de digitale arkivene som oppstår i forbindelse med slike prosjekter, åpner opp muligheten for nye måter å studere språket på fra en statistisk, språklig og litterær synsvinkel.

I 2006 innledet Nasjonalbiblioteket et digitaliseringsprosjekt med mål om å digitalisere hele sin samling av pliktavlevert og annet ervervet materiale. Samlingen består bl.a. av bøker, aviser, plakater, kart, film, musikk, TV- og radiosendinger, private arkiver etc. Det spesielle med det resulterende digitale nasjonalbiblioteket er ikke bare selve størrelsen og mangfoldet av materiale, men også det faktum at det utgjør den norske kulturarven i sin helhet.

I denne sesjonen vil fokuset være på delkorpuset som oppstår etter digitalisering av bøker og aviser, og noen av de innsikter som kan oppnås fra studier av dette korpuset (som fremdeles vokser).

For omlag 20 år siden var mangelen på materiale til å bygge nye og representative korpus en av de største utfordringene i korpus- og datalingvistikken. Et godt korpus kunne vanligvis bare lages med mye manuelt og tidkrevende arbeid. Med det digitale nasjonalbiblioteket ligger problemet heller i den store mengden og å sette sammen det riktige korpuset på basis av et sett med egnede tekster.

Fordelen med en digital biblioteksamling er at hvert enkelt element (bok, avis, film, tidsskrift etc.) bærer med seg et sett metadata i formater som gjøres tilgjengelig for søk. Dette muliggjør bygging av unike korpus, basert på et rikt utvalg av søkekriterier som kan sortere innholdet i samlingen ved hjelp av metadatatyper som publiseringsdato, materialtype, sjanger, forfatter, tema etc. Man kan f.eks. konstruere et korpus av populærlitteratur publisert i siste halvdel av det 19. århundre, og sammenligne dette med tekst fra ukeblader etc. fra 1980 til dags dato.

De tre foredragene i denne sesjonen illustrerer hvordan det digitale biblioteket kan benyttes som en kilde for språklige og litterære studier, og hvordan bibliotekmetadata kan brukes dynamisk for å bygge korpus av alle størrelser, basert på et generelt søk i bibliotekbasen for å filtrere ut nøyaktig de dataene man ønsker å analysere. Deretter kan språkdataene analyseres ved hjelp av digitale verktøy (statistisk og språklig analyse, visualisering av dataene på ulike måter, f.eks. i ordskyer, diagrammer, galakser etc.),

Ved hjelp av ren statistikk til språklig og litterær analyse ønsker vi å illustrere hvordan store digitale biblioteker kan brukes til å oppnå innsikt som ellers ville være umulig å oppnå. Vi demonstrerer også hvordan en rent statistisk tilnærming til samlingen kan brukes til å skape tjenester av interesse for forskere på ulike felt, men som også kan fungere som interessante og lærerike verktøy for skoleelever og allmennheten.

Temamodellering

Lars G. Johnsen, Arne Martinus Lindstad, Magnus Breder Birkenes (Nasjonalbiblioteket)

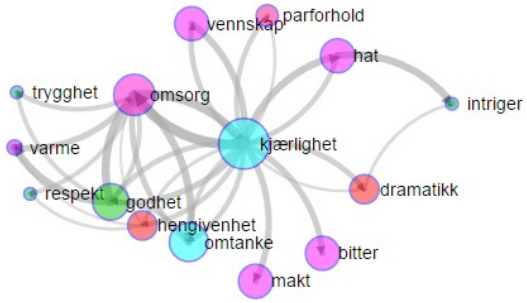
Utgangspunktet for temamodellering (engelsk topic modeling) er å betrakte tekster som ordlister. Ordlisten gis en struktur ved at den splittes opp i et sett med underlister som hver representerer et tema.

En forholdsvis ny og populær modell for temaer er den såkalte LDA-modellen (se for eksempel Jockers *Text Analysis with R for Students of Literature*, Springer 2014, og referanser der). Den krever imidlertid at antallet tema må bestemmes før analyse av en aktuell tekstsamling, samtidig som tolkningen av resultatene, i form av ordlister, kan være en utfordring, siden ordlistene må vurderes for eventuell relevans.

Vi vil presentere en metode for å finne temaer som lar teksten selv være bestemmende både for antall tema og som også gir mer støtte til tolkningen av temaene og innholdet av dem. Metoden består i å konstruere en referansemodell for temaer, som kan benyttes i analyse av tekster ved å koble tekstens innhold mot elementene i modellen.

Modellen bygges opp fra fulltekstbasen ved Nasjonalbiblioteket, der dataene hentes ut fra trigrammene i språkdatabasene.

Modellen konstrueres fra en enkelt grammatisk relasjon, det at to ord er koordinert. Strukturen bygges fra posisjonen ordene har i koordinasjoner av typen *A og B*, som for eksempel *spiser og drikker*, *rød og blå*, *kvinner og menn* og *kjærlighet og hat*. I slike tilfeller dannes en graf bestående av buer fra A til B (se figur). Grafen bygges ved å koble sammen koordinerte ord slik at A kobles til B, så B til C og kanskje også C til D.



Koordinerende konstruksjoner har et potensiale for å dekke over de fleste ordtyper, siden nærmest alle ordklasser kan koordineres innenfor klassen, som eksemplene antyder. Det betyr i praksis at man kan modellere det meste av innholdsordene i språket.

I figuren vises et tilfelle av en slik graf laget fra ordet *kjærlighet*, der det er foretatt to slike hopp: *kjærlighet og X* og så videre med ordpar som *X og Y*.

Siden alle ordene i grafene er koblet til et eller flere andre ord, direkte eller indirekte, gir det grafen en struktur. Strukturen kan undersøkes langs forskjellige akser, der de mer kjente egenskapene er sentrale element (noder med mange koblinger) i tillegg til begrepet klikker (et sett med noder der alle er direkte relatert til de andre). I grafen ser vi at *kjærlighet* er et sentralt element, og at ordene, eller nodene, *kjærlighet, omtanke* og *omsorg* utgjør en klikk av flere. Klikkstrukturen er en viktig del av grafen, og brukes til å bygge samlinger av ord som er relatert til et sentrum.

En større utgave av samme graf som vist i figuren gir blant annet følgende strukturer for *kjærlighet* (det er et titalls totalt). Klikkstrukturen er utvidet ved at klikker som deler nesten alle ord slås sammen. Mens den første er om fysisk tiltrekning, plukker den andre opp sjelelige kvaler:

Senternode: *kjærlighet*

Utvidet klikk: *seksualitet, begjær, følelser, vold, forelskelse, kjærlighet, erotikk, sex*

Senternode: *fortvilelse*

Utvidet klikk: *angst, smerte, lidelse, frykt, fortvilelse, sorg, kjærlighet, savn, lengsel, håp*

Ikke alle slike utvidede klikker har et entydig sentrum. I en tilsvarende graf for ordet *sorg* finner vi en struktur som har to senternoder:

Sentrum: *smerte, sorg*

> **Utvidet klikk:** *angst, motgang, lidelser, ensomhet, smerte, lidelse, frykt, fortvilelse, sorg, savn, lengsel, sykdom, nød*

I praktisk analyse av tekster kan man tenke seg flere måter å bruke en slik referanse på. For eksempel så kan hvert ord en støter på kobles til modellen, enten som et sentrumsord, eller som et ord i en utvidet klikk, eller som et sett av ord . Her er en setning fra "Sangen om den røde rubin", der de uthevede ordene kobles til *kjærlighet* og *fortvilelse* som sentrumsord, selv om akkurat de ordene ikke er nevnt i tekstutdraget, i tillegg til *smerte* og *sorg*:

> Hvorfor er ikke Haydn's musikk revneferdig av **sorg**, bitterhet og voldsomme **følelser**?

I foredraget vil vi videreføre en slik analysepraksis, ved at tekstens ord kobles til referansemodellen. Det som skiller metoden fra LDA (jf. Jockers) er at ordene stammer fra relasjoner mellom ord, og at ordgruppene er merket med sentrumsord. Dette er egenskaper som gjør grafene lettere å tolke.

«The Rebirth of the Author»: Eller hvordan å gjenkjenne litterær stil ved hjelp av dataanalyse

Lars Johnsen/Trond Haugen/Arne Martinus Lindstad/Magnus Breder Birkenes (Nasjonalbiblioteket)

I november 1954 utga Brann forlag i Oslo romanen *Blokk 7. Daglig liv i "Kjærlighetsredet" i en tysk konsentrasjonsleir* av det pseudonyme forfatterparet Agnes og Ernestine Lefèvre. Den ble trykt i 5000 eksemplarer hos Brødrene Jørgensen A/S, talte 255 sider, og kom i både hefta og innbundet variant, der den hefta utgaven var utstyrt med et glossy pulpcover av tegneren Birger Imerslund.

Blokk 7 utgir seg for å være de to søstrenes selvbiografiske historie om et to og et halvt år langt opphold i en tysk konsentrasjonsleir for kvinner. Det er et sted de selv omtaler som et tvangserotisk, overseksuelt inferno. Et perversitetens tempel. Historien er ført i pennen av søstrenes fetter, men det er de to kvinnene som vekselvis fører ordet i selve beretningen og beskriver den gradvise forvandlingen av sin egen seksuelle uskyld til en erfaren, depravert, masochistisk narkomani, i et monotont språk farget av nedskriverens kalde, distanserte og erotiserte voyeurisme.

Blokk 7 skapte furore i Norge da den kom. Etter noen få uker var den på førstesidene i VG, og avisene spekulerte i hvem som var forfatteren av boka. Flere mente romanen var skrevet av en annen pseudonym forfatter på Brann forlag, Janet H. Lalange, som hadde skrevet bøkene *Slavinner* og *Slavinnens seier*. 9. desember avslørte VG identiteten til mannen bak pseudonymet Janet H. Lalange, den danske reklamemannen Michael Polack.

Forfattersskapet til Blokk 7 ble aldri bekreftet bekreftet i mediene, og i Arthur Thuesens bibliografi *Beslaglagte og supprimerte bøker vedrørende Norge* nevnes ikke forfatteren ved navn.

I foredraget ønsker vi å finne ut, ved hjelp av digital humaniora, om det virkelig var Michael Pollack (aka Janet H. Lalange) som skrev Blokk 7. I hvor stor grad er en forfatters skrivemåte unik? Hvor langt kan vi gå i å fastslå at moderne digital humaniora kan vekke forfatteren til live igjen etter den såkalte *death of the author*? Hva skjer med studieobjektet litteratur, når stil blir et spørsmål om statistikk?

Vi søker svaret ved å sammenligne statistiske egenskaper ved tekstene fra et utvalg forfattere. Utvalget representerer hypoteser om hvem det er som kan ha vært opphav, og selve gjenkjenningsprosessen består i å finne ut hvem av disse som er den mest plausible kandidaten.

Forfatterens unikhet ligger i distribusjonen over ord og vendinger, og den som ligger nærmest romanen *Blokk 7* holdes opp som vinner.

For et sett med potensielle kandidater vil vi skape forfatterprofiler der disse ulike egenskaper i disse forfatternes tekster vektes opp mot hverandre. Tekstene hentes ut fra Nasjonalbibliotekets fulltekstdatabase, og kan sammenlignes med hensyn til ordvalg, typiske ord og vendinger, setningslengde, kompleksitet samt hvordan de forholder seg til hyppig brukte ord. Den siste kategorien er interessant i og med at de hundre meste frekvente ordene i norsk dekker over halvparten av all tekst som produseres. De består hovedsakelig av ord fra de lukkede ordklassene som pronomener, preposisjoner og hjelpeverb, i tillegg til punktuering som punktum og komma. For å sikre at resultatet holder så høy grad av gyldighet som mulig, lages hver forfatterprofil kun på basis av tekster av samme sjanger som Blokk 7 (roman).

Ved så å sammenlikne hver forfatterprofil med Blokk 7-teksten, vil vi kanskje kunne komme nærmere et svar på gåten om hvem som skrev den omstridte romanen.

Dagny Juel Przybyzsewska i *Samtiden*. Fjernlesning som kontekstualiseringsstrategi.

Anne Birgitte Rønning
University of Oslo
a.b.ronning@ilos.uio.no

Dagny Juel Przybyzsewska (1867–1901) har blitt viet stor oppmerksomhet i norsk kultur- og litteraturhistorie, men som bohemkvinne og muse, mer enn som selvstendig forfatter. En nyere norsk litteraturhistorie kommenterer at hennes «litterære produksjon er liten, men holdt i 1890-årene tone og trend» (Andersen, 2001: 320). Hva innebærer denne uttalelsen, og er den dekkende? Andre forskere har lagt vekt på nettopp det særegne i det lille forfatterskapet – der en kvinnestemme både bruker og vrir på tidens stemning og språk (Norseng, 1993, Rønning 2005).

For å undersøke DJP plass i samtiden, ikke som kvinne og muse i et kunstnermiljø, men som forfatterstemme, vil jeg prøve ut ulike former for *text mining* av den litterære og kulturelle konteksten for hennes utgivelser på norsk: tidsskriftet *Samtiden.* Her ble dramaet «Den sterkere» publisert i 1896 og fire kortprosatekster i 1900. *Samtiden* er digitalisert av NB, og jeg bruker ord- og frekvensanalyser og tema modelleringsverktøy som Språkbanken har lagt til rette for meg i en Jupyter Notebook. Foredraget vil relatere funnene til de nærlesningene som er gjort av DJPs forfatterskap og åpne for en diskusjon om hvorvidt fjernlesning kan være fruktbar strategi for å arbeide med kvinnelige forfatterskap som befinner seg utenfor kanon.

Referanser:

Andersen, Per Thomas, 2001. *Norsk litteraturhistorie*. Oslo: Universitetsforlaget

Norseng, Mary Kay, 1993. «Przybyszewska's Poetry in the Bright Light of Södergran's», i *Studies in German and Scandinavian Literature after 1500. A Festschrift for Goerge C. Schoolfield*, red. James A. Parente jr og Richard Erich Schade, Columbia: Camden House (s. 224–232)

Rønning, Anne Birgitte, 2002. «'Ubestemmelig utestengd utan grunn' –Noen tanker om Dagny Juel Przybyzsewskas plass i norsk litteraturvitenskap», i *Hilsen. En bok til Arne Melberg i anledning 60-årsdagen*, red. «Gruppe Oslo», Oslo: Gyldendal (s. 258–267)

Ett komplett arbetsflöde för två digitala utgåvor

Eliel Kilpelä/Johan Kylander/Anna Movall/Elisa Veit
Svenska litteratursällskapet i Finland
eliel.kilpela@sls.fi, johan.kylander@sls.fi, anna.movall@sls.fi, elisa.veit@sls.fi

Zacharias Topelius Skrifter (ZTS) och *Henry Parlands Skrifter* (HPS) är två textkritiska digitala och delvis tryckta utgåvor vid Svenska litteratursällskapet i Finland. Utgåvorna riktar sig till såväl forskare som nöjesläsare. Zacharias Topelius (1818–1898) verk omfattar prosa, lyrik, barnlitteratur, historisk-geografiska verk samt brev och dagböcker. Henry Parland (1908–1930) räknas som en av de ledande finlandssvenska modernisterna; utgåvan omfattar lyrik, prosa, essäer och brev. Redaktionen för ZTS består av 10–15 personer, de flesta historiker eller litteraturvetare. Sex delutgåvor har hittills utkommit, den första 2010. HPS är ett till omfånget mindre projekt som sedan början av 2015 sysselsätter tre redaktörer och planeras omfatta fyra till fem delutgåvor.

Vår session behandlar arbetsflödet inom projekten: hur texter, metadata och information skapas, lagras och hanteras genom hela utgivningsprocessen. Utgåvornas texter skapas som XML-filer och annotering och metadata lagras i databaser. Verktyg behövs inte enbart för att publicera texterna för slutanvändaren utan också för att underlätta det redaktionella arbetet. Vi berättar vilka tekniska lösningar som valts och ger exempel på konkreta problem och hur de lösts. Vi går igenom hur olika verktyg och arbetsprocesser hänger ihop, vad som har automatiseras och hur våra program används i praktiken. Vi presenterar tre olika infallsvinklar på arbetsflödet: en tekniskt betonad helhetsöversikt, en presentation av verktygen med betoning på funktion och användning och slutligen en konkret demonstration av en redaktörs arbete med olika verktyg.

E-redaktör Johan Kylander (ZTS/HPS) börjar sessionen med att beskriva utgivningsprocessen ur teknisk synvinkel, med fokus på de digitala verktyg som har skapats inom projekten för att underlätta arbetet. Utgivningen består av delmoment såsom skanning, optisk teckenläsning, kodning, registerskapande, granskning, variantigenkännande och publicering. Inom alla delmoment används skräddarsydda digitala verktyg som är centrala för att hantera de mängder data som uppstår. Verktygen automatiserar och snabbar upp utgivningsprocessen märkbart och skapar dessutom t.ex. automatiska register och metadata för utgåvornas sökfunktioner.

Arbetsflödet och de digitala verktygen inom två textkritiska utgivningsprojekt

Johan Kylander

Det här bidraget till sessionen *Ett komplett arbetsflöde för två digitala utgåvor* ger en översikt av arbetsflödet och de digitala verktygen för utgåvorna *Zacharias Topelius Skrifter* (ZTS) och *Henry Parlands Skrifter* (HPS).

Det första arbetsskedet för en utgåva, efter planeringsfasen, är skanning av materialet. Det resulterar i högupplösta digitala bildfiler. Om materialet är tryckt text utförs en optisk teckenläsning (OCR) för att få fram texten i digitalt textformat. Om materialet är handskrivet skrivs texten av manuellt. Både för tryckt och handskriven text skapar vi TEI XML-filer direkt. TEI är genom hela utgivningsprocessen det format som används för texterna. TEI är ett internationellt XML-baserat språk för uppmärkning av digitala texter inom humaniora. Vi har inom projekten skapat en gemensam kodningsmanual som utgående från våra textetableringsprinciper säger hur och vilka delar av TEI-standarden vi använder.

Nästa skede är att säkerställa att texterna har transkriberats korrekt. Texterna granskas två gånger. Vi har valt att för OCR-skannade texter skanna ett annat exemplar av samma text och låta ett program, Juxta, automatiskt jämföra båda texterna och markera skillnaderna. Den andra granskningen utförs helt och hållet manuellt.

Efter granskningen etableras texterna och kodningen kompletteras. Vi använder oss av s.k. ”stand-off annotation” för flera element i texterna. Det innebär att vi inte har all data inne i XML-filerna, utan att vi från XML -filerna hänvisar till externa källor. Detta gäller uppgifter om personer, platser och verk, manuskriptbeskrivningar samt punktkommentarer där utgivaren förklarar textställen. Denna information sparas i databaser. Kodningen av de semantiska elementen i texterna och deras länkningar till externa databasposter är ett tidskrävande arbete. För att spara arbetstid har vi skapat två verktyg som snabbar upp och automatiserar processerna.

TEI tagger är ett verktyg som automatiskt skapar koder runt personer och platser som nämns i texterna utgående från databaserna som skapats inom projekten. Databaserna genererar listor med olika kombinationer av för- och efternamnsfält för personer, stavningsformer, genitivformer samt förbudslistor på namn som inte skall ingå. Verktyget sorterar textsträngarna i längdordning och skapar taggar runt motsvarande partier i texterna.

Uppgifterna från databaserna kopplas ihop med de kodade textställena med hjälp av ett id. Med verktyget *TEI name selector* skapas rätt id för uppmärkta namn på personer, platser och verk i XML-filerna.

Kommenteringsarbetet inleds efter att texten är etablerad. Kommenteringen sker i en egen miljö som kallas *Kommentarverktyget*. Programmet hanterar både själva texterna och punktkommentarerna. Inom programmet kan man lätt skapa nya punktkommentarer; kommentarerna är även sökbara och kan återanvändas till flera textställen vilket är praktiskt vid t.ex. Ordförklaringar.

Olika upplagor som utkommit under författarens livstid bildar egna XML-filer, och alla textskillnader kodas. Med hjälp av ett variantkodningsverktyg som bygger på Collatex och fungerar i miljön för Existdb kan flera texter jämföras med varandra och varje skillnad i samtliga texter förses med rätt kod.

Den digitala utgåvan består av både en intern och en extern del. Den interna delen innehåller allt material från alla utgåvor som är under arbete. Den fungerar som ett arbetsområde både för redaktionen och för granskare av kommande utgåvor. Den externa delen består av de publicerade utgåvorna.

Den digitala utgåvan är kolumn- och flikbaserad. De olika flikarna innehåller utgåvans etablerade texter, de redaktionella texterna, faksimil i bildformat, de olika varianterna och utskrivna manuskript. Koden visualiseras med grafiska element och utgåvan är försedd med en sökfunktion, med Existdb som sökmotor som hanterar XML. Utgåvan, som har tagit modell av andra digitala utgåvor, bygger på PHP och Mysql och är skapad inom projektet. Materialet som visas för läsarna är TEI XML-filer som konverteras till HTML med hjälp av XSLT-stilark och Javascript.
Vi har utvecklat digitala verktyg som administrerar materialet inom den digitala utgåvan. *Publiceringsverktyget* är ett verktyg som hanterar och kompletterar XML-filer och skapar olika versioner av dem för visning på den digitala utgåvan och sökningen. Med hjälp av *Publiceringsverktyget* kan en utgivare lägga upp texten, granska den i sin slutliga miljö, ändra i

kodningen och med ett par musklick publicera om texten såväl internt som externt.
Det nyare projektet HPS bygger på liknande tekniska lösningar och arbetsflöde som ZTS. För det krävs att verktygen samt den digitala plattformen görs om så att de med några enkla parametrar kan anpassas till en annan miljö med andra databaser och texter. Våra egna verktyg och den digitala plattformen ges ut med öppen licens.

Nästan alla moment i utgivningsprocessen utnyttjar digitala verktyg. De flesta är skapade inom projektet och några har uppstått i samarbete med andra liknande projekt. Verktygen både försnabbar processen och hanterar den stora mängd data som skapas inom projekten. Utan verktygen skulle inte projekt av den här omfattningen vara möjliga.

Nästa presentation av redaktörerna Anna Movall (ZTS) och Elisa Veit (HPS) går in på hur kodningspraxisen utvecklats och anpassats samt hur de olika verktygen underlättar arbetet i praktiken. De två utgåvornas texter skiljer sig i omfång och karaktär, vilket innebär lite olika lösningar, men grundbehoven av att kunna märka upp, annotera och publicera texter är de samma. Därmed är en del verktyg gemensamma, medan andra blir meningsfulla först när materialet är tillräckligt stort. Automatisering av vissa arbetsmoment avlastar definitivt redaktören i ett jobb som ändå innebär många olika steg och stor noggrannhet.

Användning och anpassning av verktyg inom ZTS och HPS

Anna Movall/Elisa Veit

Det här bidraget till sessionen *Ett komplett arbetsflöde för två digitala utgåvor* behandlar användningen av projektens digitala verktyg och hur verktygen anpassats till de två textkritiska utgåvornas delvis olika behov. Avsikten är att gå närmare in på verktygens och de automatiserade lösningarnas möjligheter men också diskutera de åtgärder och vägval från användarens sida som de ger upphov till. Vår presentation omfattar konkreta exempel från arbetet med utgåvan.
Webbutgåvorna ZTS och HPS omfattar en rad olika texttyper: lästexter med register och kommentarer, transkriberade manuskript samt varianter. Utvecklingen av textkodningspraxis och digitala verktyg har också påverkats av att ZTS omfattar en stor textmassa i olika genrer, både skönlitterära och biografiska. HPS omfattar en betydligt mindre mängd text, men även den är fördelad på olika genrer.

XML-filer kan bilda grund för såväl den kompletta digitala utgåvan som för tryckta delutgåvor eftersom filerna kan konverteras direkt till ombrytningsprogrammet. Det här undviker problem med dubbelarbete som kan uppstå om den digitala och den tryckta utgåvan skiljs åt tidigt i arbetsprocessen. Alla texter som ingår i utgåvorna, med undantag av de redaktionella inledningarna, skapas och bearbetas i XML-format. Tryckta texter OCR-behandlas och levereras med kodning (uppmärkning) av struktur och utseende. Manuskript skrivs av och förses med manuskriptbeskrivande kodning. Sedan läggs ett lager av semantisk kodning till, vilket innebär personer, platser och litterära verk. Alla semantiskt kodade textställen kopplas till databaserna, vilket skapar person-, ort- och verkregister (som också förbättrar sökmöjligheterna) samt tooltipar i lästexten. Kodningens funktion är att skapa en textkritisk utgåva. Till skillnad från korpusprojekt kodar vi alltså inte lingvistiska element i texten.

Personregistret, tooltipar samt ett sökresultat för ”Gustav II Adolf”, som visar hur <persName>G. A.</persName> kopplats till posten för Gustav II Adolf och därför kan ge en träff.

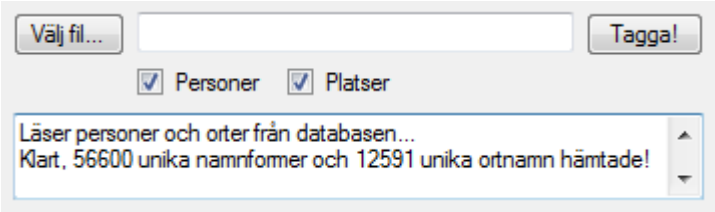
Textkodningen utgår från TEI, *the Text Encoding Initiative*s riktlinjer. Element, attribut och attributvärden måste ändå väljas ut och deras användning anpassas och avgränsas utifrån ramarna för projektet. Vid utvecklingen av kodningspraxis måste man välja tillräckligt specifik kodning som återger originalets egenskaper rätt, men även hålla uppmärkningen så enkel och intuitiv som möjligt av arbetsekonomiska skäl. Att dokumentera och följa upp är essentiellt för en enhetlig praxis.
På vissa punkter skiljer sig de två utgåvornas kodning från varandra. De två författarnas manuskript ser olika ut p.g.a. skillnader i historisk kontext, personligt temperament och syften, och kräver därför delvis olika uppmärkning. Utgåvans omfattning och texternas karaktär påverkar också behovet av semantisk kodning etc.

HTML-visningen av de kodade texterna är enhetlig inom respektive utgåva, med några genrespecifika skillnader. Det är i princip möjligt att utveckla kodningen först och slå fast visningen senare. Det är ändå lättare att balansera strävan efter att hålla kodningen så enkel som möjligt mot behovet av en tillräckligt specifik kodning om åtminstone en visningsprototyp finns tillgänglig i ett tidigt skede.

Publiceringsverktyget gör att alla ändringar i text, uppmärkning, databaskoppling och kommentarer kan synliggöras i den digitala utgåvan omedelbart. Genom den enkla publiceringsmöjligheten ser redaktörer och granskare alltid den uppdaterade versionen och kan läsa och kommentera utan att behöva använda XML-filerna.

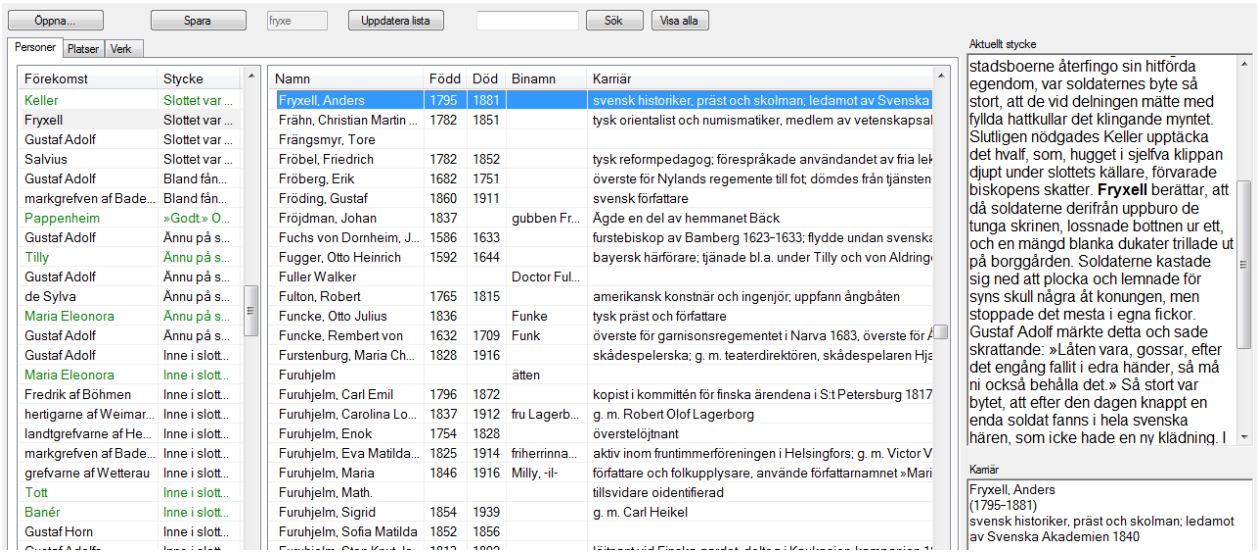
Databasernas information i kombination med kodningspraxisen gör att enahanda och tidskrävande arbetsmoment kan automatiseras i stora textmassor. För alla programmens del är redaktörens insats dock avgörande för resultatet.

Programmet *Tagger* inför person- och platskodning. Det utnyttjar databaserna för att lista namn och namnkobinationer, som jämförs med texten. Programmet strävar efter så långa träffar som möjligt, och taggar det som matchat. På så vis kodas ”Carl Magnus Rehbinder” som ett enda namn. Topelius olika stavningsvarianter finns i person- och ortdatabaserna och känns därför igen. Dessutom beaktas genitivformer. Namn som ska undantas från automatkodningen behöver dock listas för hand, och efter Taggern är en kompletterande/korrigerande kodningsgenomgång oundviklig.



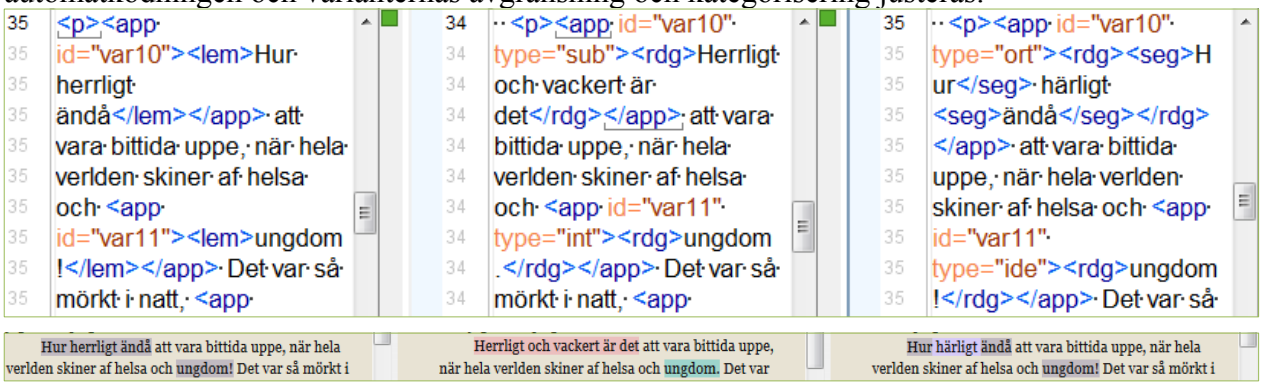
Tagger. Programmet gick igenom *Fältskärns Berättelser*, 2250 sidor, på några minuter och taggade ca 7800 textställen med <persName> /<placeName>.

Programmet *TEI Name Selector* förenklar kopplandet av semantiskt kodad text till databasposter genom att kombinera information från XML-filen och databasen. Redaktören väljer post för varje kodat namn, och programmet inför kopplingen i XML-filen, om så önskas för alla förekomster på en gång. För att Selectorn ska underlätta arbetet optimalt behöver kodningen vara väl utförd och databasposterna rätt ifyllda.



TEI Name Selector. Listan på textens personkodade ställen till vänster, varje namns kontext till höger. I mitten personposter ur databasen, nere till höger sammanfattning av den valda personposten.

Variantkodningsverktyget jämför texterna, sätter in variantkodning enligt vår praxis och kategoriserar varianterna, allt automatiskt. Variantfilerna behöver ändå bearbetas både före och efter automatkodningen och varianternas avgränsning och kategorisering justeras.



Överst variantkodning, -numrering och -kategorisering som genererats av Variantkodningsverktyget. Nederst utgåvans variantvisning, med grundtexten följd av två varianttexter. Numreringen gör att texterna scrollas parallellt, och kategoriseringen syns som olika färger (och tooltipar).

Sessionen avslutas av redaktör Eliel Kilpelä (ZTS), som presenterar det praktiska, dagliga arbetet med utgåvans verktyg. Genom en demonstration visas hur olika verktyg (den digitala utgåvans HTML-visning, Kommentarverktyget, Publiceringsverktyget, TEI Name Selector, FileMaker-databaserna och versionshanteringssystemet TortoiseSVN) används under tre arbetsmoment i slutskedet av arbetet med en delutgåva: kommentering, personkommentering med länkning samt publicering. Metoden i presentationen är att visa slutresultatet och därefter rulla upp vilka steg som var nödvändiga för att komma dit.

Ett brev annoteras – en demonstration av arbetsprocessen

Eliel Kilpelä

I mitt bidrag till sessionen *Ett komplett arbetsflöde för två digitala utgåvor* demonstrerar jag hur de olika verktygen som används inom *Zacharias Topelius Skrifter* (ZTS) används i det löpande arbetet. Som utgivare för Zacharias Topelius korrespondens har jag valt ett brev som utgångspunkten för min demonstration. Idén med demonstrationen är att visa hur redaktören använder sig av tre enkom för projektet skapade mjukvaror, två externa mjukvaror och ZTS digitala utgåva på www.topelius.fi för att genomföra tre arbetsmoment inom utgåvan: kommentering och personkommentering med länkning samt kommentarernas och lästextens publicering (för internt bruk).

De tre arbetsmomenten infaller i slutändan av arbetsprocessen med ett manuskript. De bildar s.a.s. början på slutet i en texts väg från manuskript till publicerad del i utgåvan. Grovt taget är arbetsmomenten för ett brevmanuskript inskanning och manuskriptbeskrivning, avskrift och kodning, 1:a och 2:a kollationering, införandet av semantisk kodning jämte länkning. Därtill kommer författandet av inledning och textkritisk kommentar samt intern och extern granskning. Redaktörens arbete med kommentarerna efterföljs således ännu av en del arbetsmoment, men manuskriptet har vid kommentarskedet i det stora hela fått sin digitala skepnad så som den kommer att te sig för slutanvändaren.

I min demonstration visar jag hur man inom utgåvan använder sig av de sex ovannämnda redskapen för att slutföra de tre ovannämnda arbetsmomenten. Dessa redskap är *visningen* av den digitala utgåvan, *kommentarverktyget*, *publiceringsverktyget*, *TEI Name Selector*, *FileMaker*-databasen och versionshanteringssystemet *TortoiseSVN*. De fyra förstnämnda verktygen har skräddarsytts för ZTS, det femte används med inköpt licens och det sista är en gratisvara.

Jag börjar min demonstration med att i ZTS *digitala utgåva* visa ett publicerat brev med punktkommentarer och kommentarer länkade till den semantiska kodningen, d.v.s. så som det hela visas för slutanvändaren (bild 1). Detta utgångsläge är ganska typiskt i det dagliga arbetet, och visningsläget är det jag oftast själv börjar med. I det publicerade, okommenterade brevet är det lättast att upptäcka eventuella bristfälligheter: obsolet manuskriptbeskrivande kodning, eventuella slagfel eller just personer som av en eller annan orsak blivit okodade. Mitt utgångsläge i demonstrationen är att jag upptäcker en bristfällighet: en personpost är olänkad. Dessutom vill jag förse brevet med kommentarer. Följaktligen tar jag upp de två olika typer av kommentarer jag kommer att arbeta med: punkt- och personkommentarerna, som båda visas i tooltip-rutor.

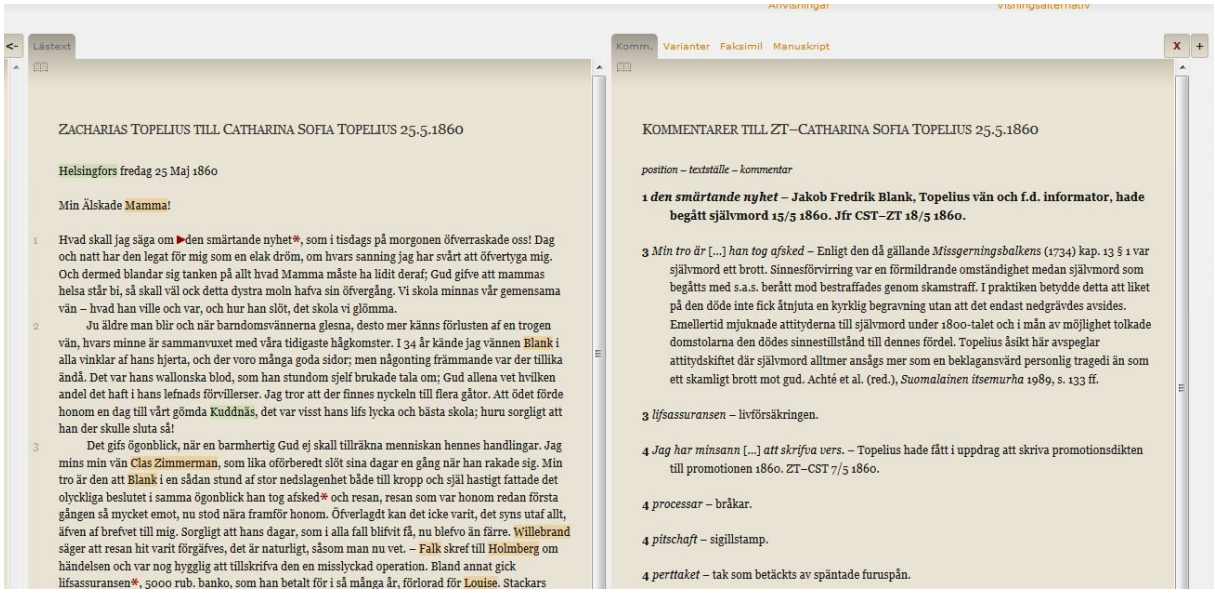


Bild 1. Den digitala utgåvan. Visningen på hemsidan.

I det följande rullar jag upp vilka steg som genomgås för att komma fram till slutresultatet. Jag börjar med att skapa en personkommentar i *databasen* (bild 2) och länka databasposten till det personkodade men olänkade textstället. Persondatabasen innehåller mer information än vad som syns i tooltipen (bild 3): upplysningar som kan bedömas vara redaktionellt intressanta införs i fälten benämnda familj, släkt, intern beskrivning, m.fl. I fälten kan man fylla i information om karriär, bekantskap med Topelius eller speciella livshändelser. Tooltipen strävar efter att vara så icke-kontextspecifik som möjligt, men eftersom persondatabasen används i hela utgåvan kan informationen i de icke-publicerade fälten visa sig vara nyttig senare i t.ex. en punktkommentar. Länkningen till databasposten går över programmet *TEI Name Selector* (bild 4). Person-, ort- och verktitelkommentarerna ersätter delvis punktkommentarerna genom att ge grundläggande upplysningar vid varje enskild förekomst. En ständigt återkommande punktkommentar skulle innebära tradiga upprepningar, särskilt som person-, ort- och verktitelförekomsterna i utgåvan är många.

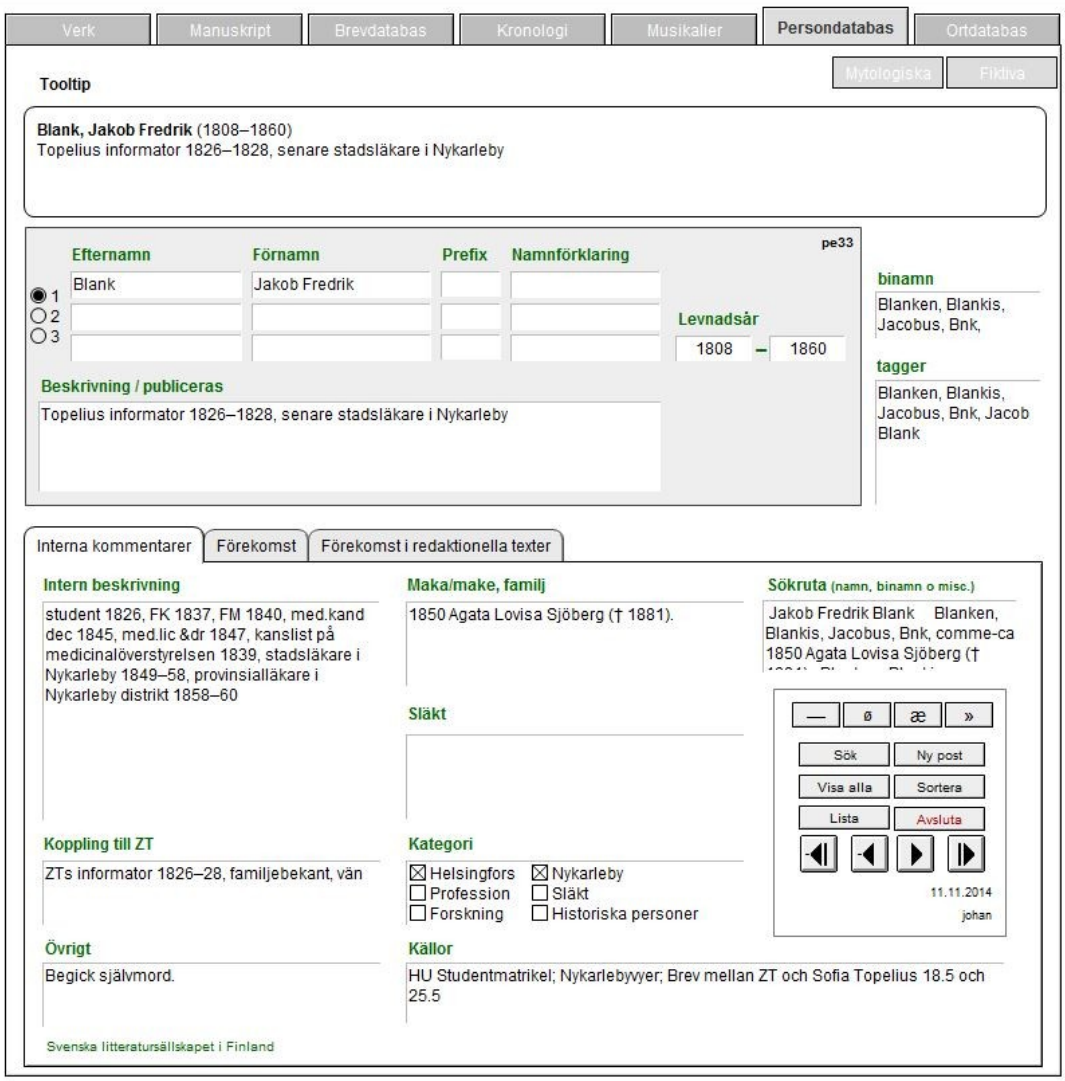


Bild 2. Persondatabasen i FileMaker.

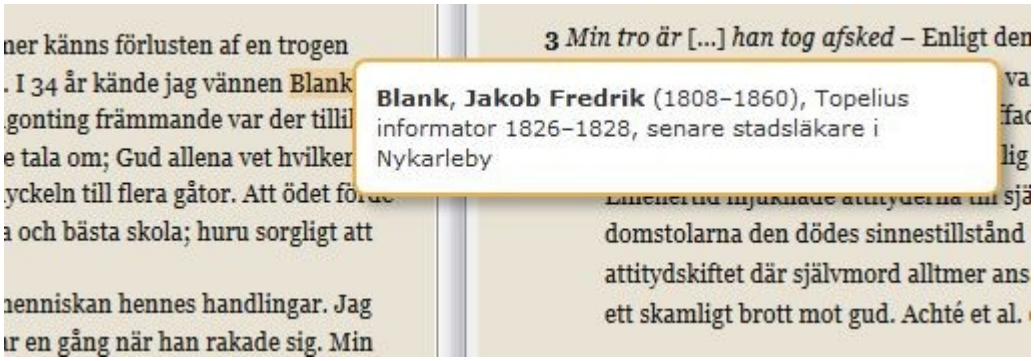


Bild 3. Persondatabasens tooltip.

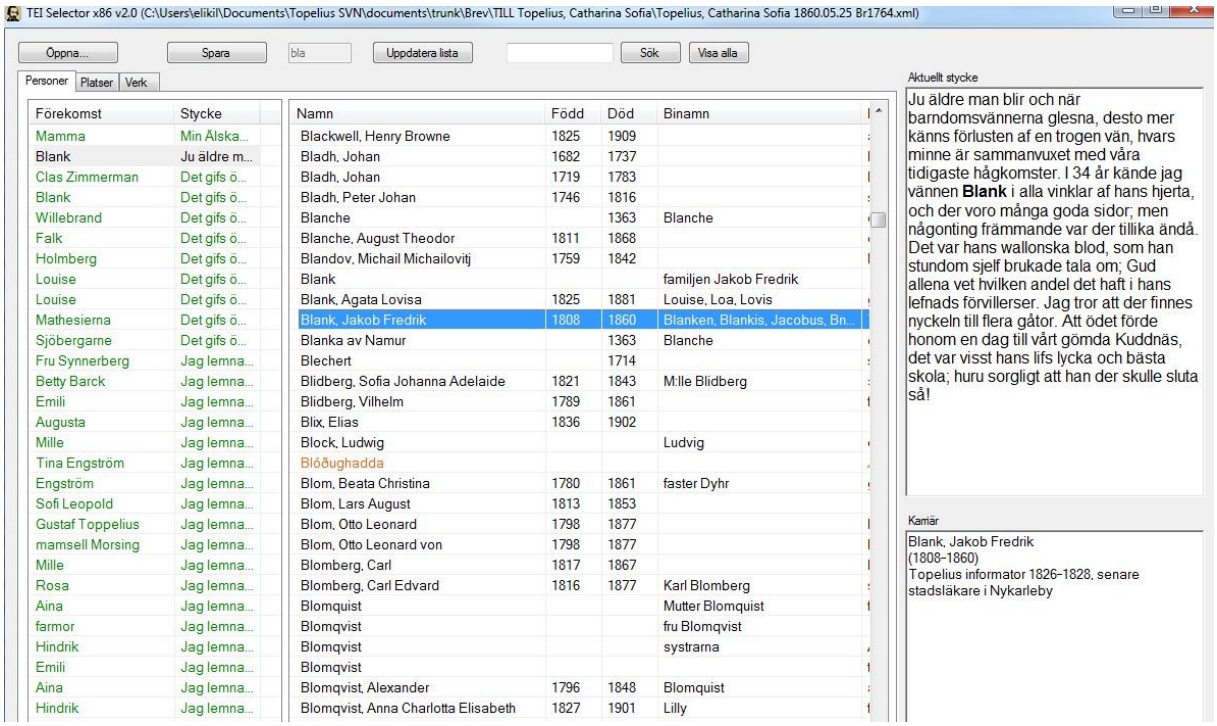


Bild 4. TEI Name Selector.

Utgåvans natur är emellertid sådan att databasernas person-, ort och verktitelkommentarer endast täcker en del av kommentarbehovet. Ett stort antal punktkommentarer är därför nödvändigt. Dessa kommentarer ger bl.a. kontextuella upplysningar om textställen, men de kan också lika bra handla om översättningar och ordförklaringar. I kommentarerna finns också länkade hänvisningar till inledningar och andra kommentarer. I demonstrationens följande steg visar jag hur en punktkommentar skapas i *kommentarverktyget* (bild 5), hur kommentaren länkas till ett visst textställe samt hur man återanvänder en kommentar på ett annat ställe. Man kan läsa texten i själva verktyget, men visningen är inte lika komplett som i den *digitala utgåvan*, och därför har dessa två verktyg skilda användningsområden för mig. I *kommentarverktyget* markerar man det textställe där man vill skapa ett lemma, varvid verktyget sätter in <anchor>-element på det stället i XML-filen.

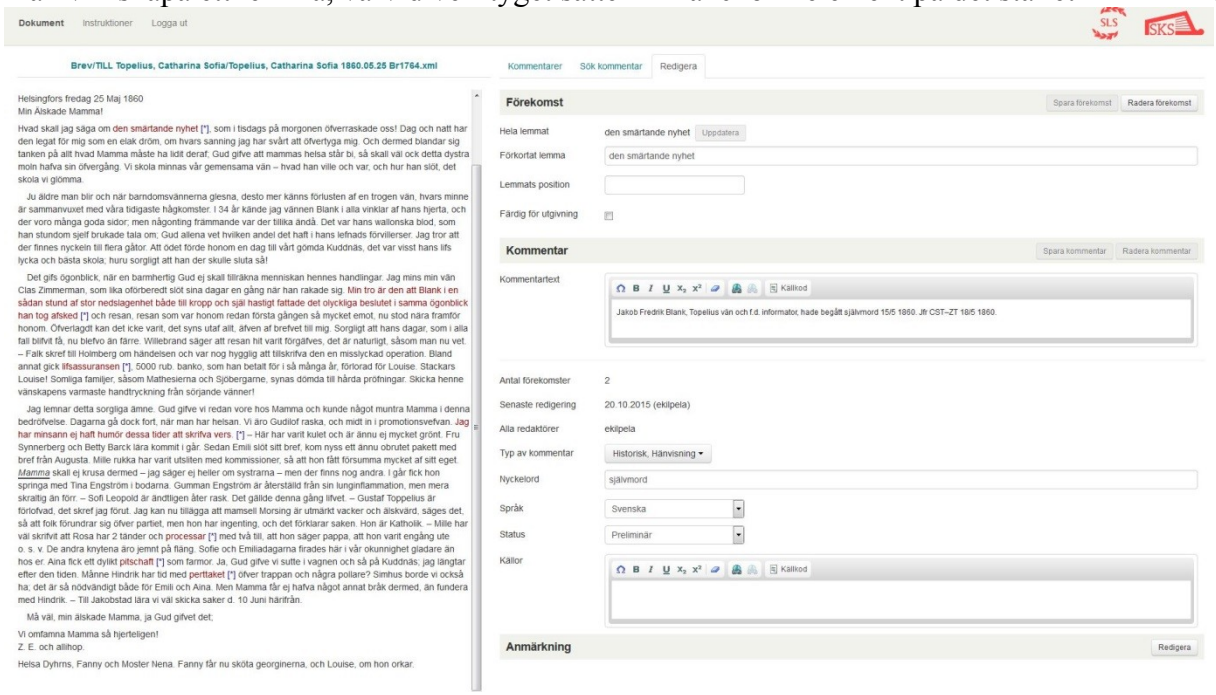


Bild 5. Kommentarverktyget.

Det tredje och sista steget i demonstrationen är att publicera de ändringar redaktören gjort i XML-filerna via verktygen. Dessa ändringar hanteras av versionshanteringsverktyget *TortoiseSVN*. För att synliggöra dem använder jag *publiceringsverktyget* (bild 6), som kommunicerar med SVN. Genom att använda *publiceringsverktyget* uppdateras det dokument som skall visas på hemsidan omgående och ändringarna som gjorts med verktygen blir synliga.

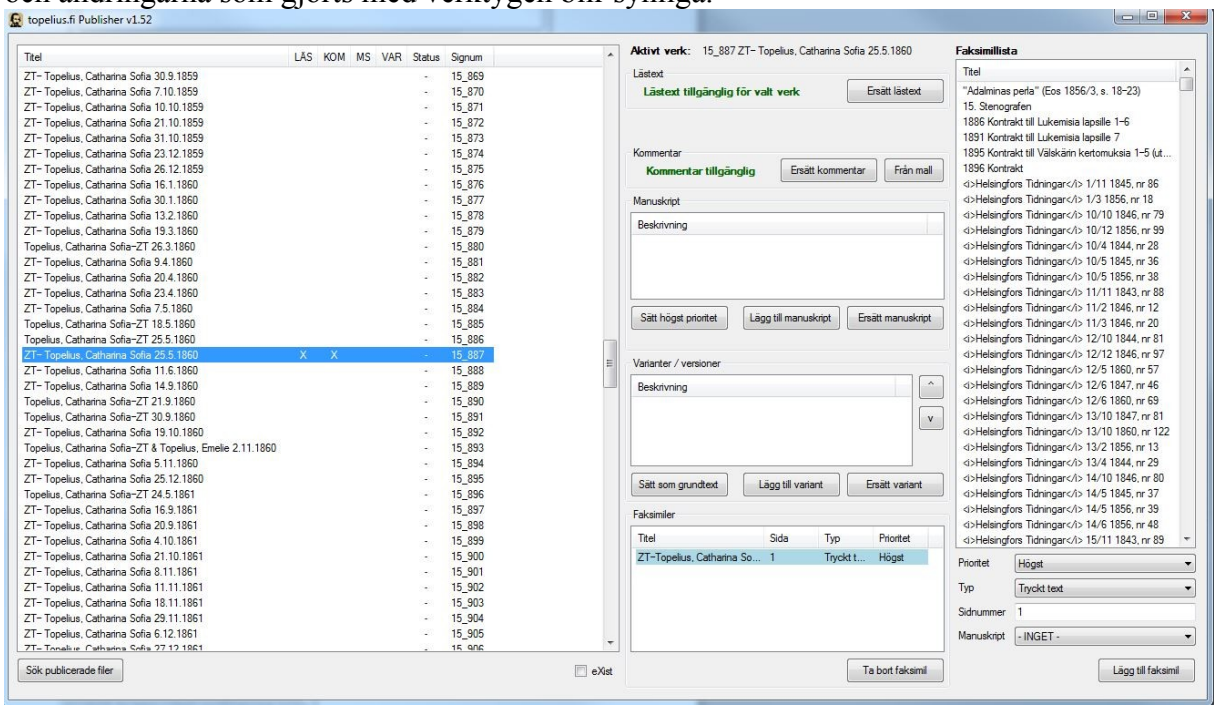


Bild 6. Publiceringsverktyget.

Spatial Humanities

Dieta Svoboda, Fabian Schwabe
University of Tübingen
dieta-frauke.svoboda@uni-tuebingen.de, fabian.schwabe@uni-tuebingen.de,

Kristina Skåden
University of Oslo
kristina.skaden@ikos.uio.no

Since the late 1980's aspects of space gained more and more importance in the humanities. In this context 'space' comprises different kinds of meaning, pertaining to geographical as well as on social definitions. Thus, activating the concept space allows broad interpretations of historical, social, and political processes. In many fields of research, the so-called 'spatial turn' supports the methodological approach of connecting textual and spatial information. Accumulations of geographic and spatial data could be understood as independent repositories for information and provide an additional category for the analysis of text. Therefore spatial data and text form two independent but equal classes of information. The combination of both will increase our understanding of topics, contents or intentions of texts.

Although spatial information is already used to visualize research topics in the humanities, the detailed analysis of geodata in the context of humanistic research projects points to a lot more options and open the possibilities to reveal hidden relations between data sets. Therefore, their future influence on the research process in humanities studies cannot be underestimated. In the last few years, the integration of spatial data in humanistic projects in general and language based project in particular can be subsumed under the term 'spatial humanities'. The aim of this subfield of digital humanities is not only to illustrate meta information within texts; the aim is also to construct new basic data by performing spatial analyses. The focus thereby lies on manual and semi-automatic text annotations, e.g. in TEI standardized XML documents that can easily be transferred to geographical information systems (GIS) for further analysis.

Researchers in digital humanities need computational assistance in order to do their research. Therefore, students have to be educated in certain computer skills. In autumn 2013 the eScience-Center of the University of Tübingen (Germany) started a teaching project in the Digital Humanities. The aim was and is to enable the students working with a computer and specific technologies. Quickly the main object and teaching field became courses in the so called spatial humanities. The benefits of the combination of textual orientated science with spatial orientated science for exploring certain research questions, became quit clear. Besides courses on technical aspects in general the eScience-Center offers courses in form of projects every term. For the teaching program 'Digital Humanities' co-seminars with other departments of the university or other institutions became important in order to realize project-based learning.

By the means of different case studies, the proposed session will not only give an introduction to the field of spatial humanities' (Bodenhamer et al. 2015; Bodenhamer et al. 2010), but will also illustrate the potential of integrating spatial data in language-based and culture-historical research and teaching projects realized by the eScience-Center of the University of Tübingen and Department of Culture Studies and Oriental Languages, University of Oslo (Norway).

Spatial Turn, Texts and Mapping – GIS in the Humanities

Dieta Svoboda

A lot of data in humanistic research include spatial information. Obviously for research areas like anthropology or archaeology that are tightly bound to time and space, spatial components are fundamental for understanding cultural phenomena or chronological developments. For these, space is just one aspect of interpreting material culture. Other fields in the humanities only started in the last decades to focus on the spatial information often hidden in their research materials. The understanding of space as an additional and inherent category of data, e.g. in the history of art, linguistics or study of literature is generally referred to with the rather abstract term 'spatial turn'.

This paper forms the introductory part of the proposed session 'Spatial Humanities'. It focuses on the leading question of the importance and possibilities of integrating spatial information in different research areas in the humanities. Three aspects will be discussed.

First, a short historical introduction to the term spatial turn will be given. It was originally established in the culture- and social science in the 1980s and the paper points to its significance for all disciplines of the humanities. The paper argues that spatial data is already integrated in many datasets.

Second, the possibilities and difficulties of integrating spatial data in the traditional analyses of data will be shown by using some example datasets from archaeological projects. The main instruments for conducting spatial analysis are geographical information systems (GIS). The wide range of different analytical features of GIS program packages can easily be linked via different mark-up languages (e.g. XML) and/or database systems to digital editing or text annotations. These approaches to texts include not only manually added detailed annotations but they also offer the opportunity to analyze large text corpora and apply the method of 'distant reading' regarding spatial information.

Last, some concluding remarks will lead to the question to what extent spatial analysis, e.g. in the studies of literature, may or may not, broaden our understanding of the intentions of writers and a general *zeitgeist* of texts which makes spatial information an important part in the interpretation of e.g. texts.

As the following papers will refer to special research projects this paper will end with some examples showing how few very basic steps of connecting text and spatial information may lead to additional benefit in the visualizations and analyses of literature. The examples originate from the teaching program of the eScience Center Tübingen and comprise different topics. They illustrate how not only the acquisition of data from scientific texts in a project that visualizes informal markets along the Silk Road, as of today, but also on the extraction of spatial data distributed in literary texts like Jules Verne's *Around the World in 80 Days*. Also, a final glimpse to historical and recent geodata and the possibilities of connecting them with other spatial information will be taken.

From Iceland to Jerusalem - the Itinerary of Nikulás of Munkaþverá as a web edition

Fabian Schwabe

The edition project of the itinerary, also called *Leiðarvísir* started as a seminar at the University of Tübingen. It is finished concerning the teaching in course, but still ongoing concerning the presentation of the results as an open website with the possibility of a theoretical everlasting annotation of the text by students or generally speaking interested people.

In a co-seminar with the chair of Scandinavian studies in winter term 2014/15 the itinerary of *Nikulás of Munkaþverá*, (d.1159/60), was annotated, edited and translated into German. The edition and translation is complete, while the annotation is done only in parts. Moreover the students worked on a transliteration of the main manuscript, it is AM 194 8vo in the Arnamagnaeanean Collection in Copenhagen (edited by Kristian Kålund 1908). Since the students had no knowledge of paleography and it was their first time working with a manuscript, their transliteration of AM 194 has to be proofread closely. This task outride of the scope of the seminar.

Working with a manuscript was not the most important part of the seminar. The focus was on an applied knowledge that could generate an annotated visualization of the described pilgrimage from Iceland to Jerusalem. Every participant of the co-seminar chose a text and wrote some annotation to certain aspects of it. All annotations are written by the students, who did the literature research, the composing of information and additional material and last but not least of course the writing of the annotation (a good overview of the current state of research gives Marani 2012). All work was peer reviewed by students of the second course on that topic in summer term 2015 and by myself.

The presentation of the work was done in Omeka/Neatline, an open source content management framework developed at the University of Virginia. It enables you to connect information, e. g. text, images etc., with a modern web map as a point, a line or a polygon. Moreover, one can include further maps as an overlay, e. g. a map of city of Jerusalem from about 1100 to 1750, and define the visibility of the geometrical forms by the use of a so called time-line and a zoom factor. In order to visualize the text itself and the annotation all information and all connections between the text and the annotations, and between the annotations they were stored in an XML file. The XML file was automatically transformed into a bundle of HTML files, which were integrated in Neatline. The HTML files are linked to each other to enable the user of the website to get a general idea of the itinerary and its meaning or the meaning of a certain passage.

The web presentation of the itinerary should appeal to non-academics, who are interested in medieval times and/or Old Norse. Step by step a reader could advance from the translation and/or normalized text towards the digital photographs of that certain manuscript. Moreover, the annotation and the embedded literature references give guidance through the text, which is far from being easily interpreted. Altogether this edition gives a new vantage point on a text, that has fascinated people fore centuries.

The itinerary is shown in standardized Old Norse and in German translation. There are more than 100 annotations created by the students. In the near future the user can also have a look on the manuscript with a transliteration and diplomatic edition of the text. As it was previously pointed out the whole project will be open for further annotation within the frameworks of university courses - not just in Tübingen – but also by interested people who would like to participate.

The Norwegian Folklore Archive and narratives of mapping.

Kristina Skåden

How shall we engage with folklore collections and historical archives in the future? In the last ten years "digital humanities" is increasingly presented as an answer to this question. A quick research shows numbers of projects, in which cultural materials were migrated into digital media. Thus, we may say with sociologist Karin Knorr Cetina's (1999) concept new 'epistemic cultures' – that in a given field make up how we know what we know, may emerge. Rather than arguing about the digital revolution capacity, this paper explores in search of new 'epistemic cultures', how digital humanities take place in actual practices, and by what means in relation to the Norwegian Folklore Archive. The key question is: How are spatial humanities and digital mapping enacting research (methods, theories, analysis) in folklore collections and archives? How are relations between the archive and the digital shaping different versions of the Norwegian Folklore Archive?

For studying these questions this paper discusses spatial humanities in relation with critical cartography and actor-network theory (ANT). Critical cartography is a theory and method that suggest investigating the role of mapping in relation to power/knowledge practices (Crampton 2010). And a central ANT-point is to understand maps as "immobile mobiles", as something that draw facts/knowledge together, are mobile, and pass on stabilized arguments (Latour 1986).

On this background, this paper further on presents a specific project on history of knowledge were Digital Humanity is applied on archive material from the *Norwegian Folklore Archive*. The project follows the map *Høiderne av de hodtil maalte Bjerge i Norske eller Rhinlandske Foder* painted (1825) by the priest Niels Hertzberg living at the Norwegian vest coast. The map has firstly close relations to the map painted by the German writer and statesman Johann Wolfgang von Goethe titled *Die Höhen der alten und neuen Welt bildlich verglichen* and secondly to the explorer and naturalist Alexander von Humboldt journey to Latin America (1799-1804). Thirdly this paper argues that this relations also relates to maps and paintings drawn by the scientist, traveler and painter Catherine Hermine Kølle (1788-1859). The aim is to discuss geospatial aspects of the archive material and secondly to discuss the benefit of turning this relations and the archive material into a digital mapping project.

Named literature

Bodenhamer, David J./Corrigan, John/Harris, Trevor M. 2010: *The spatial humanities. GIS and the future of humanities scholarship*, Spatial humanities, Bloomington.
Bodenhamer, David J./Corrigan, John/Harris, Trevor M. 2015: *Deep maps and spatial narratives. The spatial humanities*, Bloomington.
Cetina, Karin Knorr 1999: *Epistemic Cultures. How the Sciences Make Knowledge*, Harvard University Press.
Crampton, Jeremy W. 2010: Mapping: *A Critical Introduction to Cartography and GIS*, Chichester: Wiley-Blackwell.
Kålund, Kristian (Ed.) 1908: Alfræði íslensk I: AM 194 8vo, København.
Latour, Bruno 1896: «Visualisation and Cognition: Drawing Things Together» in H. Kuklick (ed.) *Knowledge and Society Studies in the Sociology of Culture Past and Present*, Jai Press vol. 6, pp. 1-40.
Marani, Tommaso 2012: *Leiðarvísir. Its Genre and Sources, with Particular Reference to the Description of Rome*, Durham.

Papers

Södra Råda: a virtual diorama in time-space

Gunnar Almevik, Jonathan Westin
University of Gothenburg
Department of Conservation
gunnar.almevik@conservation.gu.se, jonathan.westin@conservation.gu.se

In November 2001, the medieval church of Södra Råda burnt to the ground in an act of arson. The church had a log-timbered structure dating to 1309. The wooden trefoil vault had world famous medieval paintings by, among others, Master Amund. Hence, the Swedish National Heritage Board initiated a process, which is still ongoing, proclaiming that the church should be reconstructed 'as a pedagogical example to enhance craft practice and historical knowledge of medieval churches'. After several years of excavation of the remains, church grounds and cemetery, a full-scale reconstruction started in 2007. The church was supposed to be one of the best documented buildings in Sweden, however previous representations do not provide enough information about material, construction and craft procedure to move from theory to practice. The physical reconstruction involves readings of existing archive materials and archaeological record of Södra Råda and in-deep investigations of the few medieval log timber churches that still exist. The significant method for the project is the full-scale evidence based experimentations of the materials, tools and situated environments used in the process of building a log-timber church. The physical reconstruction is supposed to be completed in 2019.

Using the reconstruction of the medieval church of Södra Råda as a case, this paper presents an ongoing research project on how digital technology aimed at visualization and gaming can be used both to contextualize layers of information, and as a tool in the reconstruction process. How can we offer remote access through digital tools? How can mobile technologies and narrative devices function to communicate the historical development as well as the multi-sensory aspects of the architecture. How can we let the visitor re-enact or visually explore the underlying craft procedures of the physical reconstruction?

The research method is design driven based on workshops that bring together craftspeople, building historians and experts in different software and hardware applications. The research draws upon best practice and theoretical outcome from recent computer based visualisation projects within cultural heritage (Imitatio Maria at HumLab in Umeå, Via Appia project at Amsterdam university, Pompeji project at HumLab in Lund, CHESS project at New Acropolis Museum, MATRIX project at Michigan State University).

The results consist of a series of evidence based digital reconstructions of the medieval church of Södra Råda. Firstly, three Structure from Motion models (AgiSoft Photoscan post processed in MeshLab) from physical remains and historic photography (1908, 1959 and 1992). Secondly, a 3D model produced in animation and modelling software (Autodesk Maya) showing the different stages and transformations of the historic building through the centuries. Game engine software (Unity 3D) are used in conjunction with an Augmented Reality (AR) platform (Qualcomm Vuforia) to connect and contextualise the archive material both as a Virtual Reality (VR) environment, and as a narrative device which uses AR to convey visual information in the physical environments of the church.

During the construction phase of the physical model, the digital model is to be projected in the landscape as a reference for the craftspeople in their work, giving body to various interpretations and solutions to better inform their decision process. In the final physical model, AR will be used to complete the reconstruction by adorning the plain interior walls and ceiling with the now lost rich

decorative paintings. This is achieved by matching the Structure from Motion models processed from the archival photos with the physical reconstruction and present these as overlays to the visitors.

The results are discussed in light of the London Charter for computer based visualisation, questioning how to make uncertainties and conflicting interpretations intelligible through digital layers, and how the processes and outcomes can be properly understood and evaluated by both expert users and visitors to the heritage site.

Narrative Approaches to the Digitalization of Participatory Urban Planning:

Bringing Plot and Metaphor to PPGIS methods

Lieven Ameel
University of Tampere, Finland
lieven.ameel@staff.uta.fi

Maarit Kahila, Jenni Kuoppa, Marketta Kyttä
Aalto University, Finland

The last decades have seen a distinct "narrative" turn in urban planning practices and theory (Ameel 2016, Sandercock 2010). At the same time, planning has become increasingly reliant on digitalization in the way it carries out the participation of citizens. In planning practices, digitalization appears as a set of various instruments that can be understood as ecosystem of digital tools (Wallin et al 2010; Saad-Sulonen 2014). Amongst the most established methodologies developed to communicate with local participants and to gather information as part of participatory planning are public participation geographic information systems (PPGIS) (e.g. Brown & Kyttä 2013). These methods tend to result in a wide range of place-related information, often structured in the form of stories. The digitalization of planning processes, and the view of planning as a form of "persuasive story-telling" (Throgmorton 1996) have resulted in a number of challenges. How to aggregate the data gathered through PPGIS into meaning-making knowledge that can have effective impact on planning and policy? How to develop PPGIS that incorporate and activate story-telling mechanisms? In our paper, we will examine the potential of narrative approaches from literary and narrative studies for developing new methodological frameworks for digital participatory planning practices. The relevance of this paper lies not only in its interdisciplinarity, but also in the way it addresses key questions concerning the status of different kinds of knowledge (experiential and "soft" knowledge, in particular), as well as, more implicitly, the issues of democracy and inclusion in planning and policy.
We will focus on two specific concepts from literary and narratives studies: *plot* and *metaphor*; i.e. the causal chain of events that drives narrative, and the rhetorical tropes used to describe these changes. These concepts could further enrich the analysis and development of PPGIS in two distinct ways. First, by providing a framework with which to systematically evaluate the material gathered in PPGIS methods, drawing on a long expertise within narrative studies in analysing narrative topographies. And second, by offering new narrative approaches with which PPGIS methods could be developed in ways that strengthen the narrative characteristics of both the methods themselves, the responses given, and the way these feed into the overall planning practices in a particular project. This includes reconsidering the questions asked and responses elicited in PPGIS, as well as linking responses to broader narrative frames and the way in which metaphorical language (city as

"body"; district as "oasis") is used to describe a planning area.
We will examine narrative approaches in the context of PPGIS in two specific case study: "Enjustess" (http://www.syke.fi/projects/enjustess) and "Hanko of Memories and dreams" (http://maptionnaire.com/en/393/). The first case, which studied the use and management of aquatic environments in the Helsinki region, could be considered as a more traditional approach to PPGIS. In the case of Hanko, the traditional PPGIS was enlarged and participants were invited to provide information in a variety of forms: written texts, structured answers, and audio material (using PPGIS methods including an innovative media-installation) as well as photographs.

Sources

Ameel, Lieven (2016) (forthcoming): "Emplotting urban regeneration: Narrative strategies in the case of Kalasatama, Helsinki." In Rajaniemi, Juho (ed.) *DATUTOP*.

Attili, Giovanni (2003): "Beyond the Flatlands. Digital Ethnographies in the Planning Field." In Sandercock, Leonie & Attili, Giovanni (eds.): *Multimedia Explorations in Urban Policy and Planning: Beyond the Flatlands*. Springer, Dordrecht, 39-56.

Brown, G., & Kyttä, Marketta (2013). "Key issues and research priorities for public participation GIS (PPGIS) for land use planning: A synthesis based on empirical research." *Applied Geography 46* (1), 122-136.

Saad-Sulonen, Joanna (2014): *Combining Participations. Expanding the Locus of Participatory E-Planning by Combining Participatory Approaches in the Design of Digital Technology and in Urban Planning*. Espoo: Aalto University Publication Series.
https://aaltodoc.aalto.fi/bitstream/handle/123456789/13352/isbn9789526055435.pdf?sequence=1

Sandercock, Leonie. (2010): "From the campfire to the computer: An epistemology of multiplicity and the story turn in planning." In Sandercock Leonie & Attilli (eds.): *Multimedia Explorations in Urban Policy and Planning: Beyond the Flatlands*, Giovanni, Springer, Heidelberg, 17-37.

Throgmorton, James A. (1996): *Planning as persuasive storytelling: The rhetorical construction of Chicago's electric future*. University of Chicago Press, Chicago.

Wallin et al (2010): *Digital Tools in Participatory Planning*. Espoo: Centre for Urban and Regional Studies Publications.
https://aaltodoc.aalto.fi/bitstream/handle/123456789/4832/isbn9789526032603.pdf?sequence=1

Organizing complexity. On some challenges in the linguistic annotation of Mycenaean Greek texts.

Federico Aurora
University Library of Oslo/University of Oslo
federico.aurora@ub.uio.no

This paper presents some particular aspects of the linguistic annotation of the texts in DAMOS,

Database of Mycenaean at Oslo, [1] an annotated corpus of all the published texts of Mycenaean Greek, the only attested Greek dialect of the II millennium B.C.

Mycenaean texts are administrative documents, written mostly on clay tablets, found within the remains of palaces in Crete and mainland Greece. They amount to more than 5,900 documents, but many of them are brief or fragmentary. They are written in a syllabic script conventionally called Linear B, a script not related to the later Greek alphabets, but in scholarly practice they are conventionally transliterated into Roman characters. Although apparently an efficient tool for palace administration, Linear B is not well suited for rendering the phonetic (and consequently morphological) system of the Greek language. The language of the documents, the oldest attested Indo-European language after Hittite, has several archaic and interesting linguistic features and poses questions crucial for the history of Greek, which, especially because of the fragmentary state and the concise nature of the documents and the shortcomings of the script, are still in need of an appropriate answer. The database has been created by importing, through a PHP script, text files with updated versions of current editions into a relational database (MySql). The texts are then being linguistically annotated (partly manually, partly semi-automatically) for morphological, syntactic and lexical information. A rich set of other metadata has also been added, which is available for searches and can thus be crossed with more strictly linguistic data. This includes the epigraphical information contained in the Leiden conventions-based transliterations, which was automatically imported in dedicated tables of the relational database (one future goal is to make this data exportable in an EpiDoc compatible format).

One of the advantages of digital editions, is that they allow for a better account of the manifoldness that often lies in our data, while by use of non-digital media one needs to compromise a great deal on account of physical constraints. With 'manifoldness' I refer here especially to the presence of more possible values for a given linguistic unit (e.g. interpretative hypotheses of an ambiguous form) and of more possible representations of such a unit (e.g. different levels of normalization/lemmatization). In addition, in the case of epigraphical material and of critical editions of ancient works, multiple competing readings of one word must often be dealt with. But this, in turn, leads to more complex data sets, that pose methodological challenges regarding the theoretical and technical organization of such complexity.

All of this is well exemplified by the Mycenaean material and the way in which it is stored in our database: the data in DAMOS are stored in multiple normalization and lemmatization layers (both on the epigraphic and the linguistic side) and multiple analyses of a given linguistic unity (in turn, connected to a given reading of a word) can be entered. Thus, for example, different hypotheses on the meaning of different readings of a word (e.g. *te-qa-jo-ị* vs. *te-qa-jo*) can be entered and ranked according to different criteria (e.g. scholarly consensus). This organization of the data is, I believe, essential for work with a corpus like the Mycenaean, where interpretations are often uncertain and dependent on context and intertextual comparison. The linguistic interpretation of a given phenomenon (e.g. case syntax) can, thus, depend on competing variants of a net of hypotheses (on the number of case forms, on the phonological value of certain graphemes, etc.) and implications; it is then crucial to be able to test and compare the different possible linguistic interpretations by varying the value of certain (sets of) analyses in performing complex database queries. So, important for Mycenaen is also the possibility not only to store, like critical editions' apparatuses already do, but also to linguistically annotate different epigraphical readings, since some new data might turn up which make previously weaker hypotheses more plausible (as it is, in fact, also the case for the linguistic analyses).

1 An online version of DAMOS is since February 2013 accessible at: https://www2.hf.uio.no/damos/index/about.

Reception history across languages – a challenge for the digital humanities

Jenny Bergenmar
University of Gothenburg
jenny.bergenmar@lir.gu.se

The development within the digital humanities has changed the methodological landscape of literary historiography. Franco Moretti's call for large-scale distant readings of the transnational impact of novels, genres, and the literary markets in a European or even global perspective has been influential. Many digital humanists besides him have argued that to improve our understanding of literary epochs and trends we must move beyond minor revisions of the literary canon to include "the great unread" by means of quantitative methods. Matthew Jockers warns against drawing "conclusions about literary periods from a limited sample of texts" (2013:47), and Ted Underwood argues that "digital and quantitative methods are a valuable addition to literary study" because of their ability to "represent gradual macroscopic change" (2013:159). These scholars (and many other digital humanists) have in common that they depart from large corpuses of literary texts and use statistical analysis to make general claims about literary history, at the same time (with the exception of Moretti) delimiting their analysis to texts written in English.

This paper discusses the digital humanities methodologies applied in the writing of literary history. Since the quantitative large-scale approaches to literary history have had a strong impact on literary studies, this paper aims to 1) discuss the limitations of computer aided reading (of literary texts) as a method for historiography 2) investigate alternative digital humanities methods for the for literary historiography using other data than the literary text. Digital humanities research is stamped by cultural, linguistic and economic inequalities. A digital humanist can easily investigate a large corpus of British and/or North American literature, since it is already digitized and made available in large online collections. Moreover, the large-scale computational methods are often used for analysis of (literary) texts in printed in books, which leaves aside reception documents, for example reviews in the press, letters and other manuscripts not usually available in text format, and consequently an inconvenient material for computational analysis. In the project *Swedish Women Writers on Export in the 19th Century* we are interested in the translations, circulation and reception of texts in different language areas – objectives requiring data much less accessible.

The hegemony of the English language and the focus on text mining literary texts are two conditions in the current digital humanities research on literary history colliding with recent developments within literary history generally: the "transnational turn" (Jay 2010) and a renewed attention to reception and circulation of literature (Casanova 2004). If we ignore reception sources and draw historical conclusions from large collections of mainly literary texts, the literary history written within the domain of digital humanities will not be consistent with the recent development in literary history generally.

Yet, there are some projects that may serve as examples for this kind of research. In *Reading by Numbers* Katherine Bode sets out to rewrite Australian literary history through metadata and sales figures showing authorship, publishing, and readership circulation during certain historical periods. She calls her method a "form of distant reading", although she does not analyse a large corpus of texts using computational methods, but instead uses quantitative methods to analyse metadata and sales statistics. Bode has the advantage of a large bibliographic database, *AustLit*, where data on publications both in book format and in periodicals is aggregated. Likewise, the project *Global Literary Networks* at the University of Chicago has used metadata to map literary networks and

circulation of American modernist poetry in Japan, China and Latin America. Both these projects show the potential of using metadata for investigating the transnational circulation of literature, in the latter case also across linguistic borders. Both had the opportunity to use already existing bibliographical sources: in the former case an authoritative database with biographical and bibliographical information, in the latter case a digitized bibliography indexing 166 modernist poetry journals in Japan (Long 2015:285).

However, bibliographies such as the one used by the project *Global Literary Networks* are rare and data on the literary circulation involving more than two countries (and languages) even more so. There is a risk that the existence of large corpuses of digitised literary texts (in English and major European languages) in combination with effective text mining methods invites a certain kind of approach to literary history, using the archives with available texts and metadata instead of finding out what is lacking in these. Adeline Koh has argued that "archives are not sites of knowledge retrieval but of knowledge production" – and in fact, "monuments of states, colonies and empires" (2014:385). The digital literary canon reproduces the canonization already existing in print, which has been questioned by postcolonial, gender, and ethnic studies already in the 1980s and 1990s. This problem, Koh argues, is even more acute in the literary digital archives than in the more broadly historical ones. It is also a problem evidently more keenly experienced from the point of view of smaller languages and research concerned with transnational perspectives, for example on reception history in Europe, with its linguistic diversity and unevenly digitised literary heritage.

References:
Bode, Katherine, *Reading by Numbers. Recalibrating the Literary Field,* London: Anthem Press, 2012
Casanova, Pascale, *The World Republic of Letters,* Cambridge, Massachusetts and London: Harvard University Press, 2004
Jay, Paul, *Global Matters. The Transnational Turn in Literary Studies,* Ithaca and London: Cornell University Press, 2010
Jockers, Matthew, *Macroanalysis. Digital methods and literary history,* Urbana: University of Illinois Press, 2013
Koh, Adeline, "Inspecting the Nineteenth-century Literary Digital Archive: Omissions of Empire", *Journal of Victorian Culture*, 19.3, 2014, pp. 385–395
Long, Hoyt, "Fog and Steel. Mapping Communities of Literary Translation in an Information Age", *Journal of Japanese,* The Journal of Japanese Studies, 41. 2, 2015, pp. 281-316
Underwood, Ted, *Why Literary Periods Mattered. Historical Contrast and the Prestige of English Studies*, Stanford: Stanford University Press, 2013

Tidens Stroem – the flow of time

Temporal dynamics in N.F.S. Grundtvig's collected writings

Katrine Baunvig, Kristoffer L. Nielbo
Aarhus University
kfb@cas.au.dk

> "Since the dawn of media, we've had more on our plates than we could ever consume, but something has radically changed: never before has language had so much materiality – fluidity, plasticity, malleability […] Before digital language, words were almost always found imprisoned on a page. How different today when digitized language can be poured into any

conceivable container" (Kenneth Goldsmith, *Uncreative Writing – Managing Language in the Digital Age*, Colombia University Press, 2011, 25).

'Revenge of the text': This is the *bon mot* for Goldsmith's considerations pertaining to the shift from text materially fixated on pages, in books, on shelves, etc. into the unlimited word count in cyberspace. Vengeance - in a generously-flexible sense of the word - is also our business when undertaking the task of measuring temporal dynamics in N.F.S. Grundtvig's newly digitized collected writings. The thoughts of and the social movement associated with the 19th century religious, theological, political, historical and pedagogical thinker (1783-1872) has become a focal point in the construction of Danish national identity (John A. Hall, Ove Korsgaard, Ove Kai Pedersen (ed.),*Building the nation. N.F.S. Grundtvig and Danish National Identity*, McGill-Queen's University Press 2015). Likewise, a considerable amount of his poetical production is considered integral to Danish cultural heritage. Due to various practical reasons – one of them being Grundtvig's graphomania piling up 33.000 *published* pages - the scientific study of his writings has since the 2WW been carried out as studies of delineated periods, topics, or publications. These 'micro studies' combined with a recent post-structural trend in stressing the 'hyper-situatedness' of Grundtvig's text production (Sune Auken et al., *Ved lejlighed. Grundtvig og genrerne*, Forlaget Spring 2014) have enriched the qualitative understanding of specific genres and specific works. But this does not come without a cost: the close reading of units in the corpora longs for a complimentary distant reading of the entire body of texts. Not least because the Grundtvig collection consists of 962 publications many of which have not met scholarly interest criteria and have therefor gone under the radar. These are 'The Great Unread' calling for retribution (Franco Moretti, "The Slaughterhouse of Literature", *Distant Reading*, Verso 2013). But also missing from the (scientific) Grundtvig studies is the personal-biographical aspects. Associated with the 'hagiographical' amateur studies widespread in the late 19th and early 20th centuries, interests in temperament, psychic conditions, marital life etc. has been deemed controversial. Nevertheless we want to bring back Grundtvig to Grundtvig studies.

These ambitions are reflected in a cascading series of studies, in which we have approached the Grundtvig material from a distance, analyzing his lexical development during the 68 years of his authorship (1804-1872). Before modeling, the corpus underwent word segmentation, preprocessing and normalization. In modeling purpose, specifically numerical prediction, we have focused on sub-semantic structures such as word variance and complexity represented by measures of Type-Token-Ratio (TTR) and Entropy. Based on yearlong debates on dating the 'Kehre' in Grundtvig's work (see Kaj Thaning, *Menneske først. Grundtvigs opgør med sig selv I-II*, Gyldendal 1963; Anders Pontoppidan Thyssen, *Grundtvig og grundtvigianismen i nyt lys*, Forlaget Anis 1983), we compared several models that target a creative shift between 1824-1832. Furthermore we have found preliminary evidence in support of Hjalmar Helveg's controversial hypothesis - coined in the 1920's - of Grundtvig's bipolar condition. On this basis we want to embark on a methodological discussion promoting 'Old Criticism' (as opposed to New). In this paper we want to present our results and open a discussion of how one can pull of a computationally informed reintroduction of the biographical persona in studies in (digitized) collected writings.

Finländska klassikerbiblioteket – finländsk litteratur i det digitala landskapet

Anna Biström, Jyrki Nummi
Klassikerbiblioteket, Helsingfors universitet
anna.bistrom@helsinki.fi

I detta inlägg presenteras en digital resurs i vardande, det Finländska klassikerbiblioteket, ett samarbete mellan Finska, finskugriska och nordiska institutionen vid Helsingfors universitet, Nationalbiblioteket i Finland samt Helsingfors universitets bibliotek. I projektet bygger vi upp en omfattande och källkritisk elektronisk samling av både finskspråkig och svenskspråkig finländsk litteratur, och sammanställer forskningsdata som gäller den finländska litteraturen. Syftet är att erbjuda en kulturhistorisk resurs för forskning, utbildning och folkbildande arbete som gäller Finland. Klassikerbiblioteket som öppnade sitt gränssnitt i januari 2015, finns fritt tillgängligt på nätet och strävar efter att bli en motsvarighet till nordiska systerprojekt så som Litteraturbanken i Sverige samt Bokhylla eller Bokselskap i Norge.

Klassikerbiblioteket planeras för en bred användarkrets, såväl forskare som lärare, redaktörer och ”vanliga läsare”. Målet är att kunna erbjuda ett högklassigt elektroniskt och skönlitterärt material för den digitala humanistiska forskningen i kraftig utveckling, liksom för mera ”traditionell” forskning. Samtidigt ska materialet kunna läsas och kopieras och konverteras vidare för både forsknings- och undervisningsändamål.

Det digitala biblioteket med finländsk skönlitteratur består för nuvarande huvudsakligen av OCR-lästa PDF-filer. Metadatan följer en XML-baserad Dublin Core-standard. I framtiden planeras även publicering av verk i textformat. Även epubfiler ingår i planerna och ett mindre urval av e-böcker är redan publicerade. Korrekturläsningen av texter med en texteditor utvecklad vid Nationalbiblioteket har inletts. De korrekturlästa texterna är i formatet XML-ALTO. Alla format samlas i samband med digitaliseringen i METS-paket som lämpar sig för långtidslagring. Dessa innehåller metadata, den korrekturlästa texten och PDF-filerna.

Förutom de finländska klassikerna innehåller samlingen även andra litteraturhistoriskt intressanta verk, och i framtiden planeras även temahelheter så som finländsk barnlitteratur.
Klassikerbiblioteket erbjuder fyra olika innehållsdelar. Förutom det digitala biblioteket bygger vi upp bibliografier över forsknings- och annan litteratur kring författarna och den finländska litteraturen. Innehållet är indexerat vilket möjliggör sökningar. Därtill finns korta författarbiografier och en elektronisk referentgranskad årsbok *Joutsen-Svanen*, som fokuserar på forskning i finländsk litteratur.

I presentationen behandlas närmare projektets syften, nuläge och framtida möjligheter, frågeställningar och utmaningar.

PRINSIPALKOMPONENTER - PRINCIPAL COMPONENTS

Magnus Bognerud, Gro Benedikte Pedersen
Nasjonalmuseet for kunst, arkitektur og design
magnus.bognerud@nasjonalmuseet.no, gro.pedersen@nasjonalmuseet.no

For å lette det tidkrevende, Sisyfos-liknende arbeidet med å produsere og publisere mest mulig kunstfaglige metadata, skal prosjektet **PRINSIPALKOMPONENTER** prøve ut nye maskinelle løsninger for datahøsting og formidling.

PRINSIPALKOMPONENTER berører både konvensjonell katalogisering, formidling, bruker-generert innhold og nye fagmetoder innen maskinlæring. Vi tror maskinelle klassifiseringer kan skape nye konstellasjoner og gi ny innsikt. Bildegjenkjenning kan gi publikum nye innfallsvinkler til samlingene og samtidig generere gjenbrukbare, deskriptive metadata tilbake til museet.

Gjennom å sette prosjektet inn i en utvidet teoretisk ramme ønsker vi også å undersøke plassen maskinlæring er i ferd med å ta i samfunnet. Prosjektet grenser mellom kunsthistorie og robotikk; to fagfelt som tradisjonelt sett ligger langt fra hverandre. I dette skjæringspunktet ønsker vi å se nærmere på det som kan oppleves som urovekkende egenskaper ved algoritmens metoder. Til sammenlikning kan vi se på eksempelet hvor en robottekniker lager en menneskeliknende maskin. I en slik kontekst refererer man innenfor estetisk og psykologisk teori ofte til begrepet *the uncanny valley*[1], som uttrykker den uhyggelige erfaringen ved å se noe ukjent innta rollen som noe kjent. Maskinen er inhuman, men ser likevel levende ut ved første øyekast. For oss kan det virke som om maskinlæringsalgoritmene i større eller mindre grad beveger seg inn i dette terrenget. For vårt prosjekts del har dette ikke noe med utseende å gjøre, men med måten maskinene opererer på. Det oppstår en form for *teknisk individuering*[2] hvorpå algoritmen bruker sin menneskelige motparts handlinger *(input)* til å knytte nye koblinger og utvide sitt eget repertoar *(output)*, utilsiktet eller uventet fra programmererens side. Lag på lag med nevroner trenes opp til aktivisering når de ser spesifikke relevante trekk i bilder, og ved klassifisering blir da utallige mengder av slike mellomstrukturer kombinert for å gi ett enkelt svar. I motsetning til «perfeksjonerte» maskiner, er maskinlærde algoritmer usikre, de tviler og gjør menneskeliknende feil. Kan maskiner på denne måten bidra med metadata som åpner opp for nye innfallsvinkler til samlingene og skaper nye opplevelser i møte med publikum? Ved å undersøke dette håper vi prosjektet kan gi et faglig tilskudd i det ørkenaktige landskapet mellom fagene maskinlære og kunsthistorie, og at resultatet kan bidra til å berike, forstå og spre det viktige digitaliserings- og katalogiseringsarbeidet som museene gjør.

I forprosjektet har vi tatt utgangspunkt i et nevralt nettverk som er trent på ImageNet (skrevet i *Caffe*, utviklet av Autonomous Perception Research Lab ved Berkeley[3]). Videre har vi trent nettverket på klassifiseringer av kunstretninger, for så å legge til en visning med en t-SNE

1Begrepet er en avledning av den freudianske fornemmelsen av *Das Unheimliche,* første gang tatt i bruk av robotikkprofessoren Masahiro Mori i 1970.
2Mer om begrepet teknisk individuering her: http://www.kunstkritikk.no/artikler/mediearkeologen/?d=no
3http://bvlc.eecs.berkeley.edu/ Caffe kan sammenliknes med for eksempel Torch, utviklet av teknikere i Facebook, Twitter og Google.

algoritme[1]. T-SNE algoritmen tar mangedimensjonale systemer og reduserer dem til en todimensjonal layout. Den grupperer bilder etter motivlikhet, teknikk, komposisjon og fargebruk. Dette kan vi bruke som utgangspunkt for enkle brukergrensesnitt for sluttbrukere som navigerer i samlingene på nettet.

Vi eksperimenterer med ulike resultater av ansiktsgjenkjenning, klassifisering av kjønn, alder og stil. All informasjon henter vi fra de nettpubliserte kildene vi allerede har. I kjølvannet av vårt forrige prosjekt «Repcol» (http://bengler.no/repcol) har vi videreført samarbeidet med teknologer og designere i firmaet Bengler. Til å gjennomføre prosjektet har vi fått støtte fra Norsk kulturråds midler fra Program for digital utvikling for 2015/2016.[2]

A nation's web over time: The Development of the Danish Web 2005-2015

Niels Brügger, Janne Nielsen
Aarhus University, Denmark
nb@dac.au.dk, imvjani@dac.au.dk

Ditte Laursen
The Danish Netarchive/State Library, Denmark
dla@statsbiblioteket.dk

Traditionally, Digital Humanities has not payed much attention to born digital materials such as the web. However, during the last two decades the importance of the web has grown in most societies and it is therefore relevant to include this new digital object of study in Digital humanities.
An entire national web domain is something that we never experience as such when browsing the web, but never the less it is always there as a horizon, as the national context of our browsing. Therefore it is relevant to get as much knowledge about this as possible with a view to not only understanding the national web in its entirety, but also when analysing in-depth the web activities that take place within a nation and to which the national web constitutes the backdrop.
This presentation sets out to analyse a national web domain. The empirical basis for the presentation is the Danish national web 2005-2015 archived in Netarkivet. The Danish national web has been subject to legal deposit since 2005, with the aim to collect the entire Danish web. Our main reseach question is: What has the entire Danish web looked like in the past, and how has it developed?
First, we will discuss some general implications and methodological challenges when analysing a national web domain based on material in a national web archive. Although an entire national web domain on the online web from a given point in time is big, its mirror image in a national or transnational web archive tends to be bigger, because web archives often hold more versions of the same, even within a limited time span. The reason for this is that different archiving strategies may have been used simultaniously, or that the web crawler has followed hyperlinks to a web entity that was already archived. Therefore, with a view to getting as close as possible to studying the national web as it looked in the past when it was online and there was only one copy of each web entity at a time it is imperative to create a smaller collection from the entire web archive, in other words: a corpus.
Second, we will present a catalogue of ideas as to what could be the analytical foci to probe. This catalogue centers around these five focal points: 1) size (size of the entire web sphere, of file types, and of websites), 2) space (geographical distribution of websites), 3) structure (networks of

1Mer om t-SNE her: http://lvdmaaten.github.io/tsne/ og http://cs.stanford.edu/people/karpathy/cnnembed/
2Mer om søknaden og tildelingen her: http://www.kulturradet.no/tildelinger/prosjektmidler-museum/2015

hyperlinks), 4) vivacity (new/disappeared domain names and frequency of updating), and 5) content (file and software types, language, and semantics (e.g. word frequencies, sentiment analysis/topic modeling). We also provide some suggestions as to how the ideas can be operationalised. Finally, we will present the results of the analysis it was possible to do for the time being, namely analyses of the development of the domain names of the country code top-level domain .dk, based on lists from the national domain name registrar. The list of domain names constitutes a complete inventory of all the domain names on the national ccTLD. Therefore, it can be used to describe the development of the Danish web without looking in the archive. The following questions is analyzed: 1) What are the total number of domain names over time 2) How many domain names have disappeared and have been created compared to the previous years 3) How many domain names have changed hands compared to the previous years, and 4) How is the relationship of ownership and domains over time. We also discuss how such an analysis can be further elaborated and put into perspective by getting data from other sources.

Nordic Englishes on Twitter

Steven Coats
University of Oulu
Steven.Coats@oulu.fi

In an era when an increasing proportion of English-language communication is mediated by technology and social media services such as the micro-blogging platform Twitter, a consideration of the role of English as it continues to evolve globally must take into account local use of English in online contexts. Although studies have examined the linguistic forms, communicative functions, and discourse properties of English on Twitter (Page 2012; Bamann et al. 2014; Squires 2015), lexical and grammatical properties of Twitter English have for the most part not been considered for messages originating from non-L1 environments.

In the context of the evolving status of English as a primary global language and in light of recent studies exploring language use on Twitter worldwide (e.g. Mocanu et al. 2013), the present research takes a corpus-based approach to characterize the use of English on Twitter in the Nordic countries and shed light on some of its typological and sociolinguistic properties.

In a first step, a corpus of geo-located English-language Twitter user messages from the Nordic countries was created by accessing the Twitter Streaming API and filtering tweets using geographical coordinates; for comparative purposes, a similar corpus of messages originating from the United States was collected. After disambiguating author gender by automated methods, user messages were tagged for part-of-speech using a specialized tagger (Gimpel et al. 2011; Owoputi et al. 2013). The extent of English use was examined in light of previous findings pertaining to the geographical distribution of languages on Twitter. Frequencies of lexical, grammatical and discourse features in English-language messages were then analyzed according to geographical location and gender. Aggregate feature frequencies are interpreted as reflecting underlying communicative dimensions of English on Twitter as it is used in the Nordic countries (Biber 1995).

The research bears upon several important moments that collectively contribute to the development of global Englishes in a Nordic context. Extensive use of English on social media and Twitter in the Nordic countries has, for the most part, not been considered to represent a threat to the vitality of

the traditional languages of the region. Still, there is evidence to support the contention that English may be displacing other languages in certain functional domains in online environments such as Twitter. Secondly, the research provides a preliminary characterization of Nordic Englishes on Twitter in terms of aggregate grammatical feature frequencies, including non-standard features such as emoticons and orthographic variants. Some of the features of Nordic Twitter Englishes may represent L1 interference phenomena. A comparison with Twitter English from the United States suggests that feature frequency differences may reflect an underlying difference in communicative orientation among the Nordic Twitter userbase. Finally, the research explores the interaction of language and gender in Twitter messages by Nordic users, reinforcing some previous findings pertaining to gender and the use of language, but uncovering some unexpected patterns of use.

References

Bamann, David, Jacob Eisenstein & Tyler Schnoebelen. 2014. Gender Identity and Lexical Variation in Social Media. *Journal of Sociolinguistics* 18(2), 135–160.

Biber, Douglas. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge, UK: Cambridge University Press.

Gimpel, Kevin., Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan & Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 42–47. Stroudsburg, PA: Association for Computational Linguistics. www.ark.cs.cmu.edu/TweetNLP/gimpel+etal.acl11.pdf (accessed 30 November 2015)

Mocanu, Delia, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang & Alessandro Vespignani. 2013. The Twitter of Babel: Mapping World Languages through Microblogging Platforms. *PLoS ONE* 8(4). http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0061981 (accessed 30 November 2015)

Owoputi, Olutobi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider & Noah. A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. Proceedings of NAACL-HLT, 380–390. Stroudsburg, PA: Association for Computational Linguistics. http://www.ark.cs.cmu.edu/TweetNLP/owoputi+etal.naacl13.pdf (accessed 30 November 2015)

Page, Ruth. 2012. The linguistics of self-branding and micro-celebrity in Twitter: The role of hashtags. *Discourse & Communication* 6(2), 181 – 201.

Squires, Lauren. 2015. Twitter: Design, discourse, and implications of public text. In Alexandra Georgakopoulou & Tereza Spilioti (eds.), *The Routledge Handbook of Language and Digital Communication*, 239–256. London and New York: Routledge.

The blindspot of Digital Humanities or What Ethnography can contribute

Coppélie Cock, Anna Johansson
Umeå University
coppelie.cocq@umu.se, anna.johansson@umu.se

This presentation addresses the relation between digital humanities (DH) and digital ethnography (DE) with focus on the significance of small and thick data for understanding digital culture. Ethnography as the process and product of qualitative research is applied in several disciplines in the humanities in order to describe, explain and understand cultural practices, constructions and

interactions. Although ethnography is an important method for the study of digital culture, it is rarely addressed in literature about DH. This paper will discuss the relation between DH and DE and explain how and why DE can be a productive contribution to DH.

The boundaries of DH are continuously being formulated and negotiated, and disciplinary backgrounds matter for what scholars bring to and how they outline the field (e.g. Gold 2012). While DH takes into account the digital as both a research tool and a study object, emphasis lays to a great extent on the tool aspect and how digital technology can be used to develop new methods for retrieval, documentation and visualization of data. Typically, the field is also informed by an interest in big data and by the privileging of texts and corpora over human subjects and their practices.

In contrast to DH, ethnographic approaches to the digital (digital ethnography, online ethnography, virtual ethnography) have specialized mainly in the digital as an object of study rather than a tool. In line with the ethnographic tradition of studying people and culture, such approaches tend to focus on the practices, spaces, and socialities that emerge in relation to digital technologies. We argue in this presentation that, by opening up the field of DH to more explicitly include ethnographic perspectives and methods, we can gain a richer understanding of digital phenomena and deepen our understanding of the interplay of digital and non-digital practices and modes of expression. One of the major contributions of ethnography is the reflexive perspective on the role(s) and position(s) of the researcher in relation to the empirical material. In the paper, we will therefore discuss the researcher's role in digital ethnography and the ethical implications this position actualize in digital humanities in relation to disciplines in the humanities.

Furthermore, we want to stress that humanities, and hence also DH, should not only be about the vast amounts of information produced through big data methods, but also about particularities, in-depth analysis of small samples, and what Wang (2013), building on the concept of "thick description" developed by Geertz (1973), describes as "thick data" that reveals the social context of and connections between data points. As Wang argues, "thick data" obtained through ethnographic methods might serve to uncover the meaning of big data: while "Big Data delivers numbers; thick data delivers stories". Here, ethnographic fieldwork, such as participant observation and in-depth interviews, are some of the methods that can enable scholars in the (digital) humanities to explain and contextualize their findings as well as to better understand and bring to the fore emic perspectives, i.e. insider perspectives, self-representation and the construction of identities.

To sum up, our argument is that DH would benefit from an inclusive approach that welcomes research where the digital – and digital culture in particular – is an object of inquiry rather than a tool. As will be illustrated in this presentation, we also believe that (digital) ethnography has much to bring to the field of DH.

References

Geertz, Clifford (1973), *The interpretation of culture*. New York, Basic Books.

Gold, Matthew K (ed.) (2012), *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press.

Wang, Tricia (2013), "Big Data Needs Thick Data." *Ethnography Matters*, 13 May 2013. Accessed at: http://ethnographymatters.net/2013/05/13/big-data-needs-thick-data/.

How (not) to do textual criticism online: A user's perspective on selected Nordic digital scholarly editions

Per Pippin Aspaas & Philipp Conzett
University Library UiT The Arctic University of Norway
per.pippin.aspaas@uit.no, philipp.conzett@uit.no

Hard copies appear no longer to be the default option for scholarly editing in the Nordic countries. Online editions have have at least a place alongside more traditional 'collected works' series. In some cases, like Henrik Ibsens Skrifter, and August Strindbergs Samlade Verk, the hard copies were produced first, followed by an online edition afterwards. The same option has been chosen for the Anders Chydenius project, whereas in the case of Zacharias Topelius' Skrifter, only parts of what is published online will also be available in print. Søren Kierkegaards Skrifter have chosen to publish hard copies virtually simultaneously, while Ludvig Holbergs skrifter, Edvard Munch, Kaj Munk, and The Linnaean Correspondence, to mention only some, are so far purely online projects.

The "Nordic model" of digital humanities implies free access: the online editions are not sold, but offered freely to the public via the internet. In terms of distribution, these online editions obviously offer something that hard copies cannot: (almost) unlimited accessibility. Rarely discussed, however, are the possible downsides. Could there be that users of these online editions easily get lost? In spite of their accessibility, could it be the case that an online edition does not actually fulfil its main objective, namely providing a new framework for analyzing the work of a given creator, period etc. and to disseminate it on a hitherto unfeasible quantitative scale?

So, how can an internet site devoted to the collected works of an author or artist be designed in order to guide its user through the materials as smoothly and intuitively as possible, without lowering the standards of textual criticism? To pursue this question we have explored a number of scholarly digital editions from Norway, Sweden, Finland and Denmark, and evaluated them according to the "Criteria for Reviewing Scholarly Editions" as developed by Patrick Sahle et al. As academic end users, we have particularly been interested in aspects such as interface solutions, encoding standards, and retrievability via search engines like Google.

We name ourselves "academic end users" in full awareness of there being other and numerically more important user groups, such as pupils in schools, intellectuals, politicians or business(wo)men looking for a good quote for a speech, etc. We still maintain that it should be possible to reach out to such larger groups *without* reducing academic standards. Finally, in cases where hard copies exist alongside the online versions, we have compared the two formats in order to shed light on the question of whether the print copies serve as essential complements, or whether they are more like superfluous appendixes.

Digital Maps of Scandinavian Literature in Italy (Lt.it): a Bibliographic database of Italian translations of Scandinavian literature

Sara Culeddu
University of Florence, Istituto Italiano di Studi Germanici
sara.culeddu@email.it

I would like to share with the Association of Digital Humanities in the Nordic Countries a "work-in-progress paper" concerning the implementation of a Bibliographic database of Scandinavian literature in Italian translation.

I am working on a research focused on the literary transfer from Scandinavia to the Italian literary system; my project thus places itself within those studies working on a redefinition of the history of European and world literature based on the interaction between specific national productions and the mechanisms of global circulation; namely, the phenomenon of literary transfer (Jurt 2007). Over the last few decades, scholars like Even-Zohar, Moretti, Casanova and others have demonstrated the relevance of translated literature for the development of national literary and cultural systems, making significant progress towards a transnational literary history.

One of the main results of the project will be the implementation of a database of Italian translations of Scandinavian literature (around 5000 titles of Danish, Icelandic, Norwegian and Swedish texts). The Bibliographic database will be linked to the profiles of authors, translators, publishing houses and mediators (e.g. authors of prefaces or postfaces) in order to generate a map of the reception of Scandinavian literature in Italy through its main actors. Therefore, the creation of a digital platform will not only make available to scholars a complete bibliography of the Italian translations of Scandinavian literature published throughout the 20th Century and until today, but also give evidence to the connections among the different actors of the transfer (publishing houses, magazines, translators, consultants, agents, Italian authors, literary groups, etc.).

By creating Digital Maps of Scandinavian Literature in Italy, implemented by the DigiLab of the University of Rome Sapienza (http://digilab.uniroma1.it/), my work seeks to enhance the project History and Digital Maps of German Literature in Italy, funded by the Italian Ministry of Education, Universities and Research, that will be conducted until 2018 at the Istituto Italiano di Studi Germanici (IISG, Italian Institute of Germanic Studies) in Rome. Working on the Bibliographic database, I am currently part of a research group (FIRB) involving 3 researchers, 2 research assistants, 7 professors, and a wide network of relations that are going to be further enlarged.

The Digital Maps are hosted by an open access website made *ad hoc* and they are going to be released both as a database and a portal.

The database will include the Bibliographic data of all Italian translations of Scandinavian literature which appeared in book form between 1890 and 2015: after the closure of the FIRB project, the IISG is committed to complete and update the database.

It will be possible to consult the database through a query language (by author, by publisher, by translator, by publishing year, by publishing series etc.), to export the data in a digital format (pdf, EPUB) or to print it.

On the other hand, the portal is going to be a kind of Wikipedia for Scandinavian literature in Italy and it will contain files and prosopographical profiles about the main publishing houses and publishing series which imported Scandinavian literature in Italy, about the publishing consultants who played a major role in selecting the texts to be translated, the main translators, the main literary reviews which first introduced Scandinavian authors in Italy and so on. A system of keywords (tag) will make possible to link the files to the bibliographic data, in order to reconstruct the actual path of the texts in the Italian literary field.

The Digital Maps Lt.it is meant to be a valuable tool for Italian and International researchers in the field of Reception Studies and Translated literature, not least because it is built to be integrated with Translated literatures from other literary systems.

Transcription by dictation

Anna Clareborn, Per Cullhed
Uppsala university library
anna.clareborn@ub.uu.se, per.cullhed@ub.uu.se

Jens Edlund
KTH, Stockholm, Sweden

As digital repositories are being adapted to publish handwritten texts such as letters, diaries, notes, etc. on a large scale, more and more manuscript material will be available for research. However, to increase searchability and to be able to analyse texts by means of digital humanities tools, handwritten texts need to be transcribed. Due to the labour intensive work necessary for transcribing, handwritten texts are still a very unexploited resource within the digital humanities. There is a correlation between the speed with which manuscripts can be published on the Internet and the metadata used to describe the contents. Extensive metadata slows down the process of publishing the connected digital images, but searchability suffers. The newly launched Swedish digital repository Alvin uses a way to digitally replicate the way researchers are being presented with manuscript material in the physical library, namely in boxes with little metadata describing the contents, but instead of physical boxes in the library, they will find virtual boxes with a digital content in Alvin.

This content can be refined by various methods, for example transcription and ideally, the result of any such enrichment should be re-published alongside the original. This digital re-cycling and enrichment of handwritten texts will be described.

Transcription is done in various ways, the most common probably being plain transcription by typing the text in a word-processing software, often made by the person in need of the transcription for various research or publishing purposes. In their aim to reach better searchability and usability of their digital resources, libraries and archives have put forward handwritten texts in various crowdsourcing projects, aided by helpers that support heritage institutions by transcribing handwritten texts with various crowdsourcing tools, but always by typing.
This paper presents an alternative method for transcription, namely by dictation and post-correction as the primary way to transcribe handwritten text.

A practical example of the digital re-cycling and enrichment in the form of transcription will be

presented, namely a diary from the First World War which has been transcribed by the use of un-modified, readily available software for dictation and later post-correction. The time used in this process as compared to typing will be discussed, as well as the ergonomic aspects of dictation versus typing. The state of the art in the field of dictation software, language issues and future prospects will be touched upon.

The use of commercial software where the text output of a transcription process remains with the user and does not feed back to the software for training purposes, such as in the presented example, means that potentially valuable speech-data is lost for research and for software improvements. Ways to retain and make use of such speech technology raw-data will be presented as a potential for improvements in the field. How dictation can be used as part of university teaching and other scholarly events as a means for obtaining more speech-data for transcriptions will also be discussed. This paper combines current research in speech technology with practical suggestions for how cultural heritage institutions and universities may devise ways for a possible enrichment and re-publication of handwritten collections.

Geographical mapping of *Sagnagrunnur*, a database of published Icelandic legends

Trausti Dagsson
University of Gothenburg
traustid@gmail.com

In 1999, professor Terry Gunnell at the Icelandic University started a project to create a database of metadata about printed legends from Icelandic folk tale collections. Information for each legend in this database includes among others a bibliographic citation, summary of the story, relevant keywords related to the legend, name and home of the informant, name and home of the collector and the place names mentioned in the legend.[1]

Since 2012, the database, now containing more than 10.000 legends has been developed further with a transformation from a single table to a complex and well structured relational database where connections between legends, persons, places and keywords can be analysed. Furthermore, place names included in the database – places mentioned in the legends and homes of storytellers – have been unified and mapped geographically and to explore the data, a web-based interface has been created for researchers and others interested.[2] The interactive map can for example visualize the distribution of:

- legends told by individual informant
- legends with specific keywords
- legends from a specific collection
- homes of informants of a specific collection
- legends of a specific motif-index type (although type numbers have not been completely catalogued in the database)
- legends told by informants from a specific region

1See Gunnell, Terry. Sagnagrunnur: A New Database of Icelandic Legends in Print. *Folklore: Electronic Journal of Folklore*, 45, 151-162: http://www.folklore.ee/folklore/vol45/gunnell.pdf
2See http://sagnagrunnur.com/

Furthermore it is also possible to mix several search methods which gives a unique opportunity to look into the connections between collectors and storytellers, storytellers and keywords and interrelations between different keywords.

Folk legends are a vast resource of knowledge about the beliefs, worldview and behaviour of the 19th century and early 20th century society, especially put in a geographic perspective. Beginning with Jón Árnason in 1862–1864 and later in the nineteenth century with Ólafur Davíðsson, Sigfús Sigfússon and more collectors during the twentieth century, the Icelandic folk legends and fairy tales now fill more than nineteen collections. Among them the largest are the collections of Jón Árnason, Ólafur Davíðsson, Sigfús Sigfússon and *Gríma hin nýja*, published between 1862–1931. The distribution map can provide a macro view of the 19th century society of Iceland and micro view of lives of individual storytellers, for example the distribution of legends told by the postman Sumarliði Guðmundsson (1843-1919) who travelled regularly over long distances carrying mail and seems to have exchanged stories with local people on the way.

This project is richly inspired by the geo-semantic exploration of the legends from the danish collection of Evald Tang Kristensen conducted by Timothy Tangherlini[2] but an important difference is firstly that the material in *Sagnagrunnur* is from multiple collectors who each one had different methods of collecting, focused on different areas and talked to storytellers from different social classes. Secondly, while the landscape of Denmark is rather flat, the distribution in Iceland shows an interesting picture closely bound to the unique geography because the relativity of spatial distances is quite much if we put it in the context of the nineteenth century, when routes over mountains were much more dangerous than today, rivers were more difficult to pass and the fjords were much deeper.

Currently the database uses relational MySQL for storage and PHP for an API delivering data from the database to the interface using the JSON data format. The interface itself is browser based and is built using HTML and various JavaScript libraries. The database can be accessed at www.sagnagrunnur.com.

Project like Sagnagrunnur is an excellent example of how cultural archival data can be explored and disseminated in new ways using digital technology. In relation with other data about the history of Iceland, such as census data, genealogy data, data about individual storytellers and collectors and even spatial data from other sources this project adds a valuable layer. Furthermore, this project can be extended with the possibility of connecting the database to other databases for comparative research. This will for example be the case with a project currently in progress by the National Library of Iceland, The Arnamagnean Institute in Reykjavík and the University of Iceland about the folktale collector Jón Árnason and his social network in the mid 19th century. In that project, information about letter-writings between Jón and his collectors will be added to the database and visualized in the context of folklore collection and the role of early folklorists of the 19th century in the nation-building that was a part of the independence movement in Iceland.

In the presentation the project will be discussed in the context of folklore research and the use of geographical mapping in the field. The technology behind the database and the development process will be described and an example of the visual results and possibilities that the project brings will be showcased. Finally, the project about the collector Jón Árnason will be discussed briefly as well as other visual possibilities that utilize the database in conjunction with other

2See Broadwell, Peter M.; Tangherlini, Timothy R. *TrollFinder: Geo-Semantic Exploration of a Very Large Corpus of Danish Folklore*. In Proceedings of LREC. Istanbul, Turkey. 2012.

datasources.

Modeller av mennesker, mennesket som modell

Øyvind Eide, Oda Bakken Eide
Universität zu Köln
oe@oeide.no

Modellering som metode for forskning og utvikling i Digital humaniora har tradisjonelt vært tett knyttet til informatikk (se f.eks. Flanders og Jannidis 2015), men teoretiske arbeider har også knyttet an til andre områder så som samfunnsvitenskap og vitenskapsfilosofi (McCarty 2005 kap. 1) og semiotikk (Ciula og Eide 2015).

I humaniora har ordet „modell" vært brukt i ulike betydninger. I litteraturvitenskap har man modeller som minner om de man finner i digital humaniora, så som Propp's eventyrmodell og ulike narratologiske modeller. Ikke sjelden uttrykkes slike modeller grafisk, se f.eks. Chatman (1978 s. 267). Men bilder har også vært knyttet til modeller på andre måter, blant annet gjennom relasjonen mellom avbildet objekt og avbildning i mange typer billedkunst, inkludert fotografi. Relasjonen mellom modeller for maleri og modellering i digital humaniora har vært poengtert pedagogisk i Fishwick's innføringer i modellering; se f.eks. videoen „Model Abstraction" <http://vimeo.com/26764704> der han bruker billedkunst til å forklare modellering.

I det foreslåtte foredraget vil vi gå gjennom noen grunnleggende karakteristika ved modellering slik vi kjenner det fra digital humaniora og sammenligne med bruken av menneskelige modeller i motefotografi. Vi vil vise at det er grunnleggende likheter mellom de to betydningene av ordet „modell" og vise hvordan dette henger sammen med semiotiske aspekter ved relasjonen mellom modell og det modellerte.

Modellering i digital humaniora er en kreativ tenke- og argumentasjonsprosess der mening skapes og endres gjennom å lage og manipulere eksterne representasjoner. Hovedformålet kan være å lage et produkt, så som en vitenskaplig utgave av en tekst, eller å nå en dypere forståelse av et fenomen, som når man lager en narratologisk modell for å forstå hvordan fortellinger fungerer (Jannidis and Flanders 2013, 138; Eide 2015). Dette er et nyttig analytisk skille men må ikke oppfattes som klart: all modellering innebærer å nå ny erkjennelse og all modellering innebærer å lage eksterne representasjoner.

Når man ser på bruk av modeller i motefotografi kan man med en gang se at ett kriterium for modellering i digital humaniora ikke er oppfylt: at modellering er en forskningsstrategi. Selv om motefotografi, ikke minst den delen av bransjen som befinner seg i grenseland mot kunstfoto kan innebære en utforskning av fenomener er det ikke snakk om en forskningsstrategi i vitenskaplig forstand. Gitt denne grunnleggende forskjellen, hvilke likheter kan vi finne?

Et motefotografi innebærer en utvelgelse av aspekter ved det avfotograferte, som alt fotografi gjør. Men utvalget skjer i form av klare og veldefinerte kriterier der målet er å uttrykke et helt bestemt budskap gjennom bruk av menneskelig avbildningsmateriale. Dette skjer fra start til slutt i prosessen, fra valg av person til å bli avfotografert via make-up og styling til lyssetting under fotografering. Prosessen fortsetter etter at bildene er tatt gjennom utvalg, redigering, og kontekstualisering. Det er ikke nødvendigvis slik at alle aspekter ved modellens personlighet

fjernes, men målet er å ta med de aspektene som er nødvendige for å formidle budskapet, og å legge til andre aspekter gjennom prosessen. Der et familiebilde har som hovedformål å presentere en person gjennom en forenklet form har et motefotografi som hovedformål å presentere et kommersielt budskap der visse aspekter ved personen som er avbildet er brukt for å skape dette budskapet.

Selv om formålet med motefotografi ikke er forskning finner vi andre grunnleggende likheter med modellering i digital humaniora. Man skaper en ekstern representasjon – et bilde i digital form og ofte også uttrykt i andre fysiske former så som papir. Den eksterne representasjonen har som formål å etablere mening, dels gjennom en aktiv kreativ prosess hos fotografen og miljøet rundt ham eller henne, dels gjennom en aktiv tenkeprosess hos mottakeren. Denne tenkeprosessen er ideelt sett (sett fra avsenderens ståsted) styrt av bildets retorikk men opplevd som funnet ut selv.

Basert på denne sammenligningen vil vi bruke det kommersielle motefotografiet til å belyse det retoriske og potensielt manipulerende ved bruk av modeller i digital humaniora.

Referanser

Chatman, S. (1978). *Story and Discourse: Narrative Structure in Fiction and Film.* Ithaca: Cornell University Press.

Ciula, A. and Eide, Ø. (2015). „Modelling in Digital Humanities: Signs in Context." Digital Humanities, Sydney, Australia, June 29 June–July 3 2015. URL: http://www.oeide.no/research/DH2015/CiulaEideAbstractDH2015.pdf

Eide, Ø. (2015). „Ontologies, Data Modeling, and TEI." *Journal of the Text Encoding Initiative,* 8.

Fishwick, P. „Model Abstraction." [Video presentation]. Duration: 53:15. URL: http://vimeo.com/26764704

Flanders, J. and Jannidis, F. (2015). „Knowledge Organization and Data Modeling in the Humanities." [White Paper]. 2015. URL: https://opus.bibliothek.uni-wuerzburg.de/frontdoor/index/index/docId/11127

Jannidis, F. and Flanders, J. (2013). „A Concept of Data Modeling for the Humanities." Digital Humanities, Lincoln, Nebraska, USA, 16-19 July 2013: 237–39. URL: http://dh2013.unl.edu/abstracts/ab-313.html

McCarty, W. (2005). *Humanities computing.* Basingstoke: Palgrave Macmillan.

Streaming Cultural Heritage: Following Files in Digital Music Distribution

Maria Eriksson
Department of Culture and Media Studies, Umeå University
maria.c.eriksson@umu.se

With over 75 million active users, the Swedish streaming service Spotify has re-written the map for how music travels between producers and consumers, and rapidly grown into one of the world's largest and fastest growing music services. This paper presents an overview and some preliminary results from a 4-year long interdisciplinary research project that studies cultural, political, economical, and historical dimensions of streaming in general, and Spotify in particular. Placed at the intersection of media studies, social anthropology, ethnology, economic history, computer science, and film studies, the project's five researchers engage in interventionist methods that trace the 'social life' and 'cultural biographies' of streamed music files, rather then those making, selling or collecting it. Weaving together fields such as media archaeology, software studies, critical media industry studies and the digital humanities, the research is based on a series of sub-project experiments that attempt to reverse engineer Spotify's algorithms, aggregation procedures, metadata structures, and valuation strategies for music. The project involves various kinds of "bot experiments" that explores, mimics and subverts Spotify's notions of usage and listening, as well as the establishment of a record label for research purposes (something which involves strategic productions and releases of music on the Spotify platform). Through such experiments we ask: How is streamed music commodified and valued? What normative worldviews are promoted and materialized by streaming architectures? What sounds are (or aren't) perceived as music by Spotify and adjacent services, and how is streaming re-designing social contracts within the music industries?

For more information, visit: www.streamingheritage.se

Historians digging in the text mine

Exploring blended close & distant reading of technical journals to understand Finnish history of industrialization, 1880-1910

Mats Fridlund
Aalto University, Finland
mats.fridlund@aalto.fi

Petri Paju
University of Turku, Finland
petpaju@utu.fi

The paper describes an explorative project to use various text mining tools to confirm and challenge existing knowledge about the history of technology and industrialization in general and that of Finnish history of industrialization in particular. Since 2012 one of us (Fridlund) have done preliminary research to explore the possibility and fruitfulness of using various text mining methods

within history of technology, so far with a focus on the use of digital humanities methods – primarily textual topic modeling and spatial information methods and visualizations. This has primarily consisted of meta methodological research and development studying how such methods have previously been used in historical studies. However, recentlöy we have actually conducted a first practical project testing the feasibility for historical research on Finnish industrialization. This first practical pilot project used digitized Swedish-language Finnish engineering journal from 1880-1910 that has been digitized by the National Library of Finland and made openly available online. All the issues for the journals has not been completely digitized for the period but fruitful enough to be used to give a preliminary assessment of the feasibility for further historical research. The journals studied were the primary technical journals during the period in Finland in Swedish and Finnish. The larger future study want to explore various transnational issues of technoscientific knowledge and technology transfer of the period, especially what kind of foreign technologies that were of interests to the engineering community and how that changed over time during the period – a rising interest in electrotechnology and waning of steam technology perhaps – as well as what countries and places that were seen as influential and in the focus of the engineering community and how that changed over time – declining interest in English and Russian technological development and rising interest of German and US would be a hypothesis.

Part of these issues were explored during the Spring in a first small student-driven project. The occasion for this first pilot project was our participation in a hackathon in May 2015 were humanities and computer science students were brought together for 5 intensive days to work together on a practical digital humanities research project and where we supplied the possible topics and historical context and suggestions for feasible research questions, the National Library of Finland provided digitized corpus text material (engineering journals). The student group that consisted of a mix of undergraduate and graduate students with competence in humanities, computational linguistics and computer science chose to focus on the introduction of electric light to Finland at the late 19th century and early 20th century as a case study. Their project combined close and distant reading through using various text mining and data visualization methods that was scaffolded by humanistic close-reading interpretation of selected texts.

Through this we got a first rough proof-of-concept pilot project of the ground that explored how historical data mining of engineering journals could confirm existing research results as well as reveal new possible historical conceptualization about the transnational technological aspects of Finnish industrialization 1880-1910. The results will be further described and discussed in the presentation.

Integrating heterogeneous digital text resources to analyse complex neologisms in Norwegian

Anje Müller Gjesdal
Norwegian School of Economics
anje.gjesdal@nhh.no

Gunn Inger Lyse
University of Bergen
gunn.lyse@uib.no

Recent years have seen an increase in research on dynamic language phenomena within the disciplines of linguistics and discourse analysis, especially related to societal issues, such as climate

change (see e.g. Tvinnereim & Fløttum 2015). At the same time, research infrastructures such as META-NET (www.meta-net.eu) and CLARIN (www.clarin.eu) work to document language resources and tools (LRT) and to streamline data formats to facilitate the combination of resources in research. The Norwegian contribution to the European research infrastructure for digital linguistic data, CLARIN, is emerging through the national project CLARINO. Research infrastructures such as CLARINO may enable researchers to combine the use of different resources efficiently, allowing for new perspectives on established as well as emerging research methods and questions, opening up, for instance, for a fruitful collaboration between linguistics and discourse analysis on the one hand, and data sciences on the other. However, an effort is needed to provide researchers with a toolbox of data sets and tools that are user-friendly enough to take full advantage of the potential offered by recent technological developments.

In tandem with academic infrastructure initiatives, we observe the emergence of linguistic data resources that were not originally developed for language research purposes, yet reveal an interesting potential for linguistic research, particularly with respect to hypothesis generation. Relevant examples in a Norwegian context are databases such as *Atekst* and *Holder de ord?* [‘Do they keep their word?’]. *Holder de ord?* is based on the proceedings of parliamentary debates, which has been made publicly available as structured data in XML format, while *Atekst* is a privately owned digital news archive.

This paper presents a use case where separate digital resources, each developed for different purposes, are used for a linguistic analysis of an emerging environmental neologism, *det grønne skiftet*. By studying this expression across independent resources we measure both its frequency as well as its spread across corpora and domains, which we hypothesize to be an indication of its neology status (see Gjesdal & Lyse 2016). Observations of the expression *det grønne skiftet* in non-academic text sources (*Atekst* and *Holder de ord*) confirmed our initial hypothesis that the use of this expression is indeed increasing in Norwegian. Unfortunately, the query possibilities in these resources are extremely limited for scientific purposes. For instance, *Holder de ord?* does allow queries using wildcards (a search such as *grøn* skift** comprises both the word form *grønt* and *grønne*, and *skifte*|*skiftet*), but it is not possible to express that the wildcard should stop at word spaces. Thus, a search for *grøn* skift** will return any any match where the string grøn* precedes *skift**, even if there are five sentences between the two words. A better-suited corpus is therefore arguably the Norwegian Newspaper Corpus (NNC) (Andersen & Hofland 2012), a monitor corpus of Norwegian news text, developed by Uni Research Computing. The NNC is available via the CLARINO infrastructure through the *Corpuscle* corpus management and analysis system, which has been developed for annotated corpora (see http://clarino.uib.no/korpuskel/). It is an interesting corpus because it is updated with new material every day, and it covers a range of content domains, spanning from political discussions to sports, from 1998 until today. NNC seems to confirm the observation from Atekst and *Holder de ord?* that the expression has seen a conspicuous growth in the news media.

In conclusion, the paper shows an increase in the expression, but importantly the material analysed provides insufficient for a conclusion at this stage. We therefore argue that improved access to research corpora, as well as corpus analysis platforms with adapted, easily accessible and user-friendly tools to analyse such data, would significantly advance the analyses of dynamic language phenomena that require access to heterogeneous types of digital language data. We also argue that infrastructures such as CLARINO have a key role to play in this development.

References

Andersen, Gisle (ed.) (2012): *Exploring Newspaper Language: Using the Web to Create and*

Investigate a Large Corpus of Modern Norwegian. Philadelphia: John Benjamins Publishing Company.
Gjesdal, Anje Müller & Kjersti Fløttum (2014): "Conceptualisations de la notion du futur dans les blogs portant sur les changements de climat." *Revue française de linguistique appliquée* 19.1: 33-48.
Gjesdal, Anje Müller & Gunn Inger Lyse (forthcoming, 2016). "Exploring an environmental neologism in Norwegian across corpora". *Neologica*.
Hinrichs, Erhard & Krauwer, Steven (2014): "The CLARIN Research Infrastructure: Resources and Tools for E-Humanities Scholars." In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), May 2014, 1525–31.
Kristiansen, Marita (2012): "Using web-based corpora to find Norwegian specialised neologies." *Communication & Language at Work* 1.1 : 11-20.
Lyse, Gunn Inger & Gisle Andersen (2012): "Collocations and statistical analysis of n-grams." In. Andersen, Gisle (ed.). *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian* (2012): 79.
Renouf, Antoinette (2014): "Neology: from word to register". In. Vandelanotte, L. et al. (eds.) *Recent Advances in Corpus Linguistics. Developing and Exploiting Corpora.* Amsterdam: Rodopi.
Tvinnereim, Endre, and Kjersti Fløttum. "Explaining topic prevalence in answers to open-ended survey questions about climate change." *Nature Climate Change* 5.8 (2015): 744-747.

Counterfactual history in simulation video games: a methodological approach to studying historical consciousness

Ylva Grufstedt,
Department of Philosophy, History, Culture and Art, University of Helsinki.
ylva.grufstedt@helsinki.fi

As part of my PhD on historical video games, I study how player-created counterfactual scenarios (or "what if"-scenarios) may be utilized as a key methodological concept for the study of historical consciousness. The overall purpose of the study is to better explain and understand the processes through which historical knowledge and ideas about the past are communicated today – processes known as historical consciousness and historical culture (Karlsson & Zander 2009, 37–49). Historical video games are part of such communicational processes in that they at times exhibit content with explicit references to the past. They display, highlight or utilize "history" (herein understood as eras, processes, events or people in the past) in order to create a player experience. Under certain conditions, such video games also allow players to create alternate historical scenarios – counterfactual history. Given video games' interactive features, players are faced with historiographic agency, which may result in counterfactual scenarios – in other words, expressions of historical experimentation and exploration.

Theoretical debates on definitions of historical consciousness are so far plentiful, yet empirical studies on the topic, particularly with regards to games and play, are lacking. One of the major reasons appears to be the absence of applicable, straight-forward methods. In this abstract, I theorize about the potential benefits of developing a methodology based on historical consciousness

at the creation of counterfactual scenarios in video games. I propose working towards a methodology (among others) that focuses on the interplay of historical consciousness and counterfactual scenarios generated by players in historical simulation video games. I hope to build on this theoretical approach and do a reception study on players of historical simulation games (currently a work in progress).

Subscribing to Jörn Rüsen's definition, historical consciousness is the ability to logically explain the present in light of the past through three competences: experience, interpretation and orientation. The competences correspond to three elements which constitute historical narratives: content, form and function (Rüsen 2006, 69–70). Historical simulation video games invite players to evaluate and experiment with historical factors and events in a manner comparable to said competences, and with similar elements, by manipulating the game according to a fixed set of rules and game mechanics. Such simulation gameplay creates historical scenarios within the games. The study of historical consciousness does not require said historical scenarios to align with historical actuality – focus should be on understanding how historical content is managed or perhaps rearranged by the player. The dissonant nature of counterfactual scenarios might give insights into the player's notions on causality, continuity, change and other historiographical factors (Chapman 2013, 223-231), and furthermore, what this might imply in terms of historical consciousness.

While video games provide some freedom to "change the course of history", it is important to note that games are digital products based on systems which frame and limit the players' agency. Furthermore, video games are products of historical culture and their content is not all-encompassing, but based on developer choice. While this limitation likely has bearing on the results of the study, and will be taken into close consideration, it is worth noting that systematic rulesets may also work in favor of the researcher, by their way of mapping the historical scenarios, and other data, consistently. Naturally, it also allows for larger amount of collected data, as well as inviting researchers interested in historical culture and consciousness to further analyze collected expressions of historical consciousness using computational tools.

In summary, historical simulation video games provide an interactive platform for creating counterfactual scenarios of the past. Given the theoretical correlation between competences of historical consciousness and historiographic gameplay, counterfactual scenarios provide a key point of interplay. In light of this interplay, I propose that working towards a methodological approach using player created counterfactual scenarios for the study of historical consciousness, is a path worth exploring.

References:

Rüsen, Jörn, "Historical Consciousness: Narrative, Structure, Moral Function, and Ontogenetic Development": Peter Seixas (ed.), *Theorizing Historical Consciousness*, University of Toronto Press: Toronto (2006), p. 63–85.

Karlsson, Klas-Göran, "Historiedidaktik: begrepp, teori och analys": Karlsson, Klas-Göran & Zander, Ulf (ed.), *Historien är nu–En introduktion till historiedidaktiken*, Studentlitteratur: Lund (2009), p. 23–70.

Chapman, Adam, *The Great Game of History: An analytical approach to and Analysis of the Videogame as a Historical Form.* PhD thesis. University of Hull (2013).

Using electronic maps to show speech data

Oddrun Grønvik, Jon Holmen
University of Oslo
oddrun.gronvik@iln.uio.no, jon.holmen@iln.uio.no

Using digital maps as a demonstration and access point to localised research data is an old idea which until now has been little explored, owing to (1) the need for electronic speed and capacity to upload maps, and (2) the fact that good electronic map data have been hard to find at an affordable price for digital humanities. Since 1.1.2014 the public map agencies of the Nordic countries give free access to state-owned quality controlled and very detailed map data. An electronic base map has been made showing Norway with of local administration borders ("kommunegrenser") from the 1940-s, at its most fine grained. In 2014, a number of speech data collections at the Norwegian language collections at the University of Oslo where therefore connected to this electronic base map.

The Language Collections accumulated at the University of Oslo (1885 - 2015) encompass dialect materials from the whole country. They include (1) a collection of ca 650 paper maps (*Norsk dialektatlas*) made at UiO, showing language conditions and language change (created 1940-1970) (2) a catalogue set (*Målføresynopsen*) showing ca 4000 word forms and morphological forms from all local administrative areas, (3) a large number of dialect word collections (incorporated in the general resource *Setelarkivet* or digitised individually in *Ordbokshotellet*). These resources are used as raw materials for the dictionary database *Norsk Ordbok*, enabling the use of maps to give geographical information about the forms and senses of dictionary head words.

In order to achieve maximum usefulness for these data, a model is needed which allows discrete use of separate categories of data, and free combinations of different categories and instances of categories. For instance, all maps use word forms which also are headwords in official orthography. Maps should therefore be searchable by standard word forms, inflected forms and dialect forms. Data presented in digital maps also need to be searchable for place (ordered hierarchically from "local administrative area" to "region") and for other metadata, in as uniform a format as possible. This can be done by analysing map content and giving each information item (from the particular to the more general) a separate layer. Users of the map database will then be able to select (a) the resource(s) to draw map data from, (b) one or more data categories, (c) data organisation, (d) data presentation in the map (symbols, colours etc).

The data tested on the current map solutions are (1) entry information in *Norsk Ordbok* (location of forms, definitions and usage examples), (2) *Målføresynopsisen*, (3) *Norsk Dialektatlas* and (4) *Ordbokshotellet.*

URLs:
Norsk Ordbok: http://norskordbok.uio.no
Målføresynopsisen: http://www.edd.uio.no/synops/work/hovedside.html
Norsk Dialektatlas, Setelarkivet, Ordbokshotellet: http://www.edd.uio.no/perl/search/search.cgi

Perceptions of Truth/Untruth in Ibsen Criticism

Annika Rockenberger, Jens-Morten Hanssen
University of Oslo
annika.rockenberger@ilos.uio.no, j.m.hanssen@ibsen.uio.no

In works of fiction, there is no such thing as truth. In literary criticism, however, there could be – and often is. Critics claim that a literary work (f.i. *Pillars of Society*) is a 'true' depiction of contemporary life or that a literary character (f.i. Hedda) possesses no 'truth value' or that the ending of a literary work (f.i. *A Doll's House*) is 'untruthful'. Critics use this notion in ways quite different from everyday usage, deprived of its connection to the realm of logic and objective facts. Throughout Ibsen's life, his works were a site of contestation over aesthetic, literary, artistic and intellectual matters, cultural policy, and political ideology. Claims of truth/untruth are one of the strongest rhetorical devices used in these contestations.

In this explorative paper we will interrogate the rhetoric of truth/untruth as employed by critics in 600 book and theatre reviews of Ibsen's works from the period 1850-1906. The reviews stem from newspapers and journals in Norway, Denmark, Sweden, Germany, England and France, allowing us to look into the critical reception of Ibsen's works across national borders and from a multilingual perspective. The reviews have been digitized as part of the Ibsen.nb.no website project at the National Library of Norway. The multilanguage corpus of reviews has been made available as .html-files that we converted into .txt-files for further exploring them with tools for computer-aided text analysis. The provided meta data (date, publisher, publication, author, title of reviewed work) has been extracted seperately and stored in a bibliographical database (Zotero). The corpus of reviews has then been divided by language into 5 subcorpora (Dano-Norwegian, Swedish, English, German, French).

For the analysis, we will, firstly, extract key words as well as their collocations and their frequencies over time with Voyant tools (voyant-tools.org & beta.voyant-tools.org) for each subcorpus and for the corpus as a whole. The focus will expecially be on notions of 'truth/untruth' in their different linguistic forms: truth/untruth, true/untrue, truthful/untruthful, Wahrheit/Unwahrheit, wahr/unwahr, wahrhaft/unwahrhaft, sandhed/usandhed, sand/usand, sandfærdig/usandfærdig, sanning/osanning, sann/osann, sannfärdig/osannfärdig, vérité/contre-vérité, vrai/faux, etc. We will then look at the underlying semantic structures (topic modeling) with MALLET (http://mallet.cs.umass.edu) as well as Papermachines (http://mallet.cs.umass.edu) and relate them to Ibsen's works, to the critics, the journals and newspapers, and the different nations to search for patterns of correlation and co-occurance and thus hope to shed some light on the critical reception of Ibsen during his lifetime across national and linguistic borders.

Digitala lärobjekt för det omvända klassrummet inom humaniora

Mika Hietanen
Uppsala University
mika.hietanen@littvet.uu.se

Med hjälp av det omvända klassrummets pedagogik och digitala verktyg kan man frigöra både lärar- och studentresurser och därmed effektivera undervisningen. Akademisk humaniora ställer

dock särskilda krav på såväl digitala verktyg som på det omvända klassrummets övningsmoment. Bägge kräver tillämpningar som skiljer sig från andra discipliner. Vad gäller de digitala utmaningarna har vi å ena sidan studenternas något sämre beredskap att hantera program och digitala miljöer. Å lärarens sida är det inte lika lätt som t.ex. inom data-, system- och naturvetenskaper att finna material som kan presenteras grafiskt, t.ex. genom tabeller eller schemata eller som processer. På motsvarande sätt arbetar man inom humaniora inte på samma sätt och i lika hög grad med konkreta fallstudier eller processer med ett klart facit. Följaktligen innebär en digitalisering och omvändning av kurser inom akademiska ämnen såsom historia, litteratur och retorik ett annorlunda grepp.

Under höstterminen 2015 har inom Uppsala universitet det omvända klassrummets pedagogik tillsammans med interaktiva digitala lärobjekt piloterats inom det humanistisk-samhällsvetenskapliga vetenskapsområdet. Utgående från ett av dessa projekt, som gäller Retorik A, behandlas implikationer för universitetsundervisning, med målet att effektivera lärprocessen. Projektet har följts av pedagogiska utvecklare och studenterna har vid många tillfällen gett återkoppling. Även undervisningen är väl dokumenterad.

Vad gäller det omvända klassrummets pedagogik och bruket av digitala lärobjekt visar resultatet på tydliga vinster för både lärare och studenter. Studenterna lyfter i återkopplingen särskilt fram två aspekter. Å ena sidan flexibiliteten när det gäller det digitala förhandsmaterialet som uppfattas såsom stimulerande och en god grund för seminarierna. Olika lösningar för att åskådliggöra ett teoretiskt och i hög grad diskursivt material, har varit lyckade. Å andra sidan beskrivs seminarierna som mera givande än tidigare och det lyfts fram att studenterna fått en möjlighet att engagera sig både som individer och i grupp. Från lärarens sida sett har engagemanget hos studenterna varit klart större och kursinnehållet har kunnat diskuteras och övas mera utförligt än vanligt i och med att studenterna kommit bättre förberedda till seminarierna. De få negativa kommentarerna gäller smärre inledande tekniska svårigheter.

Det omvända klassrummet finns väl beskrivet tidigare. Här ligger fokus på samspelet mellan de digitala verktygen, studenter och lärare, samt dessas roll för seminarieundervisningen i samband med en teoretisk kurs inom humaniora på universitetsnivå. Här diskuteras olika typer av uppgifter som kan användas för att täcka de särskilda behov som gäller inom ett vetenskapsområde där texter, deras tolkning och förståelse, spelar en central roll. Här diskuteras även hur en videoföreläsning kan skapas från material som traditionellt ansetts vara lämpligast för bokform.

De tekniska redskapen har för detta projekt bestått av *Apple Keynote* och *Quicktime* för produktion av bildspel och en ljudutrustning för inspelning av narration. Videorna har integrerats i verktyget *ScalableLearning* som bland annat möjliggör ett pedagogiskt arbetsflöde, interaktiva uppgifter, diskussion online, samt uppföljning av studenternas progression och eventuella svårigheter i arbete med materialet. Uppgifterna och hur dessa fungerat beskrivs närmare och kontrasteras med andra former av undervisning.

Ibsen and the Digital Humanities

Frode Helland
Centre for Ibsen Studies, University of Oslo
frode.helland@ibsen.uio.no

Julie Holledge
Flinders University, Adelaide, South Australia/Centre for Ibsen Studies, University of Oslo
julie.holledge@flinders.edu.au

This paper will summarise the digital humanities methodology and the research findings from an eight-year collaborative research project that has addressed a simple question: what accounts for the global success of *Et dukkehjem*? The results of this project will be published as a book length study by Palgrave Macmillian in 2016.

Cultural transmission and *adaptation* are the twin themes of the study and distant visions are the methodological starting point that unite them. The literary historian, Franco Moretti is central to the approach: he notes that "distance is however not an obstacle, but a specific form of knowledge: fewer elements, hence a sharper sense of their overall interconnection. Shapes, relations, structures. Forms. Models" (2005, 1). Looking from a distance at 3787 production records of *Et dukkehjem* in the IbsenStage database, this project has used the patterns in data visualisations to identify new sites for investigation. These distant vision have led to the close examination of aspects of the production history of the play that suggest new ways of explaining its global success.

The presentation will begin by introducing IbsenStage and the visualization techniques used to display the *Et dukkehjem* dataset. This brief introduction will be followed by two examples of data interrogations that have contributed to our understanding of the performance life of Ibsen's most popular play.

The first example will demonstrate how the application of network analysis to 825 Nordic production events has uncovered an astounding degree of interconnection. The analysis of these interconnections has lead to the identification of unbroken lines of artistic transmission from the first performance of *Et dukkehjem* in Copenhagen in 1879 to major productions of the play presented in Norway in the 1990s.

The second example will show how mapping trajectories of artists associated with performance of the play in more than three countries has revealed a complex pattern of theatrical touring responsible for the worldwide transmission of the play.

The Creation and Features of the Online Archive Devoted to Sigurður Guðmundsson, the Painter, and the "Evening Society"

Olga Holownia
University of Iceland
olga@hi.is

"Inventing Culture: Defining Sources of Theory and Inspiration and the Long-term Results of Culture Creation by Icelandic Intellectuals 1857–1874" is an interdisciplinary project which brought together specialists from a variety of fields, including folkloristics, literature, theatre and museum studies, history and design. Carried out between 2011 and 2014, it resulted in the creation of an extensive archive that maps out the intellectual history and the widespread cultural influence of the Icelandic "Evening Society" (Kveldfélagið) and one of its founding members, the artist and designer Sigurður Guðmundsson (1833–1874). "Inventing Culture" was conceived as a project which aimed to collect as much available material as possible relating to the work of Sigurður Guðmundsson and his colleagues over the space of a very active 17 years in the cultural history of Iceland, spanning the period from the publication of a polemical article on a traditional dress to his death in 1874. The project, led by Terry Gunnell (University of Iceland) and Karl Aspelund (University of Rhode Island and University of Iceland), supported by the Icelandic Research Fund (RANNIS), was a unique collaboration between Icelandic scholars at home and abroad and key Icelandic archive institutions: the National and University Library of Iceland (Landsbókasafn Íslands – Háskólabókasafn), the National Museum (Þjóðminjasafn Íslands), the Árni Magnússon Institute for Icelandic Studies (Stofnun Árna Magnússonar í íslenskum fræðum), and the National Archive of Iceland (Þjóðskjalasafn Íslands). One of the main objectives of the project was to graft the Sigurður Guðmundsson archive into the network of existing digital resources available both in Iceland and abroad.

The aim of this paper is to present the features and functionality of the digital archive produced by the project (www.sigurdurmalari.hi.is). Its core collection comprises the transcripts of the Society's Minutes, letters between various members of the Society and Sigurður Guðmundsson as well as an extensive selection of his essays, drafts of his speeches and other informal writings. The transcripts are accompanied by digitized manuscripts. This collection is complemented with image galleries including Guðmundsson's published drawings, designs of stage backdrops as well as unpublished sketches. The open-ended structure of the archive means that it can be used to engage both scholars and students in close reading and textual interpretation, building on their new media literacies. Equally, the vast corpus of textual and visual material allows future researchers to look at anything from a development of a particular concept to cultural correspondences between Iceland and the rest of Europe during the period of forty years covered by the archive.

Apart from being a useful research tool for scholars and an educational "sandbox" for students, the archive has already provided a deeper understanding of "cultural planning" (Even-Zohar 2008) while its format has brought to light networks, associations and friendships which determined the ultimate effectiveness of the Evening Society's culture-shaping initiatives. Furthermore, the research conducted by the project investigators (see, for example, Aspelund 2011, Gunnell 2012) has underscored one of its assumptions about how important the international context of romantic nationalism was for the development of Icelandic culture at this time. Indeed, the timeline which is an integral part of the archive, places it firmly within that context. Finally, the paper will also

discuss the theoretical and practical considerations that pertain to a large digital humanities project of this kind, placing it within the broader landscape of the development of Digital Humanities in Iceland.

References:
Aspelund, K. (2011). Who controls culture? Power, craft and gender in the creation of Icelandic women's national dress. Unpublished PhD thesis: Boston University, Anthropology and Material Culture.

Even-Zohar, I. (2008). "Culture planning, cohesion, and the making and maintenance of entities." In Beyond descriptive translation studies: Investigations in homage to Gideon Toury, (pp. 277–292). Anthony Pym, Miriam Shlesinger & Daniel Simeoni (Editors). Amsterdam & Philadelphia: John Benjamin PC.

Gunnell, T. (2012). National folklore, national drama and the creation of visual national identity: the case of Jón Árnason, Sigurður Guðmundsson and Indriði Einarsson in Iceland. In Timothy Baycroft and David Hopkin (Editors), Folklore and nationalism in Europe during the long nineteenth century, (pp. 301–323). Leiden, Boston: Brill.

The Musical Gestures Toolbox – Towards Digital Musicology in Matlab

Bo Zhou, Alexander Refsum Jensenius
University of Oslo
a.r.jensenius@imv.uio.no

This paper presents our work on the development of a toolbox for Matlab entitled "Musical Gestures Toolbox". This toolbox is aimed at solving pressing needs for the analysis of multimodal data in music research, or more specifically in the field of cognitive musicology. Here the overarching aim is to understand more about how people experience music, often using observation studies of people's behaviour to/with music as the point of departure. Such studies are often based on a combination of time-based audio, video and sensor data, typically combined with qualitative interviews and interpretative data. The challenge, then, is to develop analysis tools that allow for handling and working with such media and data in a systematic manner.

We have chosen Matlab as the development platform since it is readily available in our community, and there are already several pre-existing toolboxes to build on. This includes the "Mocap Toolbox" made for the analysis and visualization of music-related motion capture data. The "Music Information Retrieval (MIR) Toolbox" is another relevant toolbox, which is aimed for the extraction of musical features from audio data and the investigation of relationships between sound and music features.

While the two above mentioned toolboxes are useful for studying motion capture data and audio, respectively, they are very differently designed and it is not possible to make combined analyses of audio and motion capture data. Furthermore, there is no integration with video analysis (computer vision). We have previously developed a toolbox for music-related video analysis in the graphical programming environment Max, with a number of novel visualization techniques (motiongrams, motion history images, etc.). These techniques are commonly used in music research, but are not currently available in Matlab, hence our interest in also porting these tools.

The new Musical Gestures Toolbox for Matlab aims to be a swizz-army knife for music researchers. It builds on the above mentioned toolboxes, and also adds new functionality. The following main structures are included:

- tools for importing and exporting media and data
- transformation and data processing tools (trimming, cropping, rotating, etc.)
- visualisation tools (motiongram, spectrogram, mocapgram, etc.)
- mid and higher level feature extraction tools (quantity of motion, sound level, etc.)

Future additions will include more advanced machine learning techniques for feature extraction and classification across the different data types and media.

Dealings with Uneven Corpus – Experiences from the Use of a Difficult Research Data

Antti Kanner
University of Helsinki
antti.kanner@helsinki.fi

This paper examines technical challenges inside historical newspaper corpora I have encountered while conducting research on patterns of historical meaning change. I will present solutions to these obstacles, and also offer evaluation of the methods chosen. The technical difficulties covered include 1) problems with imperfect Optical Character Recognition (OCR), 2) morphological analysers failing due to the research data differing from what the analyser is based or trained on, and 3) overall historical processes projecting heterogeneity to the research data on many levels. As forms of meaning changes are common topics in not only linguistics but also historical studies and social sciences, and as historical corpora tend to suffer from similar inadequacies, the findings should also have more general relevance.

Perceptions and solutions proposed in this paper originate as by-products of my doctoral research in the field of Finnish linguistics, which deals with historical meaning change. As my research data I use the collection of Finnish newspapers digitized by the National Library of Finland. The collection contains all the newspapers published in Finland in 1822-1910. The collection is however very heterogenous, as the press and other forms printed public discourse in Finnish only developed in Finland during the 19th century. Historical variation in conventions of typesetting, editing and orthography as well as paper quality used for printing make it very difficult for OCR systems to recognize characters with 100 percent accuracy. Kettunen et. al. estimated that OCR accuracy is actually somewhere between 60 and 80 percent. To tackle this problem, I devised a Python script that calculated the most probable candidates for original inputs of the misinterpreted OCR output based on a word frequency list and a training data of manually corrected words.

The OCR problems connect my research to previous work done on normalization of historical corpus data and my approach is similar is similar to ones proposed by e.g. Baron & Rayson (2008) and Bollman (2012). However, not all problems in the automatic recognition of the data come from OCR problems or even historical spelling variation. Much is also due to actual linguistic factors: the 19th century saw large scale dialectal, orthographical and lexical variation in written Finnish. To

exemplify the scale of variation, when a morphological analyser for Modern Finnish (OMORFI, Pirinen 2015) was used, it could only parse around 60 percent of a 19th century word from the Corpus of Early Modern Finnish (CEMF). This problem is also identified by Bollman (2013: 60-66). My response was to modify the morphological analyser for the dialectal and historical variation. Since my research deals not with the actual task of normalization of historical orthography, but with word frequency distributions calculated from the historical data, I have been more interested in the effect the nature of the data has in these distributions, than in the correction rate of automated normalization. The use of evaluation methods reflects this focus: I have used series of manually corrected samples to ones in the raw data and ones produced by the corrections routines I have used and tried to estimate the behaviour of word frequency distributions, such as type counts of derivative and morphological categories and the share of hapax legomena.

Finally, as said, the collection itself is uneven due to the historical processes that produced it. For example, word counts of yearly subcorpora rise following a power curve from 77 thousand in 1822 to some 261 million in 1910, with outliers produced by historical discontinuities such as changes in censorship legislation. Due to this, as well as inaccuracies necessarily persisting after the automatic corrections, we are still far from the point where results from the data could be used without reservations. I will also present some simple computational approaches I have used to take account these reservations in my research.

Baron, A. & Rayson, P. (2008) *VARD 2: A tool for dealing with spelling variation in historical corpora.* In Proceedings of the Postgraduate Conference in Corpus Linguistics.

Bollmann, M. (2013). *Automatic Normalization for Linguistic Annotation of Historical Language Data.* Bochumer Linguistische Arbeitsberichte: 13.

Bollmann, M. (2012). *(Semi-)Automatic Normalization of Historical Texts using Distance Measures and the Norma tool.* In: Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2), pp. 3–14. Lisbon, Portugal.

CEMF. *Corpus of Early Modern Finnish*. (http://kaino.kotus.fi/korpus/1800/meta/1800_coll_rdf.xml)

Kettunen, K., Honkela, T., Lindén, K., Kauppinen, P., Pääkkönen, T. & Kervinen, J. (2014). *Analyzing and Improving the Quality of a Historical News Collection using Language Technology and Statistical Machine Learning Methods,* in IFLA World Library and Information Congress Proceedings : 80th IFLA General Conference and Assembly. Lyon. France. (http://www.ifla.org/files/assets/newspapers/Geneva_2014/s6-honkela-en.pdf)

NLDC. *National Library's Digital Collections*. (http://digi.kansalliskirjasto.fi/?language=en)

Pirinen, T. 2015. *Omorfi—Free and open source morphological lexical database for Finnish*, in Proceedings of the 20th Nordic Conference of Computational Linguistics NODALIDA 2015 (http://www.computing.dcu.ie/~tpirinen/Pirinen-2015-nodalida-omorfi.pdf)

The Many Shapes of Vladimir Nabokov's 'Lolita' and 'Mademoiselle O': a Quantitative Analysis

Eugenia Kelbert
University of Passau, Germany
zhenia.kel@gmail.com

Sasha Rudan
University of Oslo
sasharu@ifi.uio.no

This paper, a meta-study of cross-lingual stylistic comparison, seeks practical solutions to quantifying those of the possible stylistic markers that current language processing tools are already capable of tracking. Is style separate from the linguistic norms of a given language? Can quantitative tools help up visualize selected features of a text as a multidimensional space and thus compare the stylistic profile of different texts across languages, with reference to balanced corpora in each language? This study draws on an original system developed by the authors, Bukvik, and aims to make a contribution to the subfield in Digital Humanities known as computational stylistics or stylometry.

On the example of Vladimir Nabokov's English, Russian and French texts, with a focus on his novel "Lolita" and his short story "Mademoiselle 'O'" (both texts exist in multiple versions in two or three languages), the paper demonstrates the ability of digital methods to inform close reading of large and small texts through focusing the reader on salient patterns observed through quantifiable stylistic parameters.

The method of textual analysis presented in this paper is based on an original method of visualization of individual words in a literary text as a network, as a 'Society of Words'. This further allows for isolating semantic strands within that network with the help of Part-of-Speech tagging and WordNet. Just like in actual reading, Society of Words gives us the ability to focus on thematic or grammatical aspects of the novel that interest us the most. This can be characters (NER), thematic clusters (constructed manually or with the help of Wordnet), parts of speech etc. In each case, the network allow us to see which words are associated – form a series throughout the work – with the given node representing, say, Humbert Humbert in "Lolita", a given group of words (such as colours) or a part of speech (such as adjectives). The basic connection is essentially the pattern whereby two words occur in close proximity of each other and are associated with each other repeatedly throughout the text. In addition, this approach allows for an added ability to probe deeper into the textual patterns than we otherwise could, for example, through applying quantitative methods to the network repeatedly as the network grows and changes following the novel's development, and focusing on a differential comparison that can then reveal those changes. Another way is through the concept of shared connections or shared neighbourhoods, common in SNA research but never before applied to literature. This allows us to consider also pairs of words that, while not occurring together, share multiple words and contexts that are associated with both of them. In this way, as we can see, words and concepts can be associated one with another in a tangible way that affects interpretation without ever occurring themselves in the same sentence.

Other features of Bukvik complement this approach and render it more complex. This includes a multidimensional lexical analysis of Nabokov's vocabulary in English, Russian and French. Put

together, the methods presented in the paper are used to complement traditional close reading as the numbers and diagrams are 'translated back' to the passages in the text they represent. This work relies on interdisciplinary collaboration between a literary scholar and a team of professional programmers. The tool's modular structure ensures its relevance beyond the set of features existing tools are capable of tracking today and extends the relevance of the stylistic profile model beyond the specifics of the material considered in this paper.

Exploring Scholarly Craftsmanship and Digital Research Methods in nodegoat

Pim van Bree, Geert Kessels.
LAB1100.
pim@lab1100.com, geert@lab1100.com

In this paper we will explore methodological approaches to digital research practices in the humanities by means of the web-based research environment nodegoat (http://nodegoat.net/). For the majority of researchers in the humanities, automated research processes are unattainable as their data may be dispersed, heterogeneous, incomplete or only available in an analogue format. For an art historian studying local urban networks, no resources exist that will be suitable for any form of automated analysis. A practice that is far more suitable in this respect is the old fashioned card catalogue system, in which all relevant objects and their varying attributes and relations are described. Since the emergence of digital research tools, scholars in the humanities have the ability to create digital card catalogue systems (databases). Although a number of historical databases exist, few individual researchers produce a database as an integral part of their individual research practice.

We have developed nodegoat to facilitate humanities researchers in creating and managing datasets in a dynamic environment with a focus on relationality and diachronic, geospatial, and social contextualisation.

nodegoat is a web-based research environment that facilitates an object-oriented form of data management with an integrated support for diachronic and spatial modes of analysis. This research environment has been designed to allow scholars to determine and design custom relational database models. nodegoat dynamically combines functionalities of a database management system (e.g. Access/FileMaker) with visualisation possibilities (e.g. Gephi or Palladio) and extends these functionalities (e.g. with in-text referencing) in one web-based GUI. As a result, nodegoat offers researchers an environment that seamlessly combines data management functionalities with the ability analyse and visualise data. The explorative nature of nodegoat allows researchers to trailblaze through data; instead of working with static 'pushes' – or exports – of data, data is dynamically 'pulled' within its context each time a query is fired. The environment can be used in self defined collaborative configurations with varying clearance levels for different groups of users.

nodegoat follows an object-oriented approach throughout its core functionalities. Borrowing from actor-network theory this means that people, events, artefacts, and sources are treated as equal: objects, and hierarchy depends solely on the composition of the network: relations. This object-oriented approach advocates the self-identification of individual objects and maps the correlation of

objects within the collective.[1]

In the case of a research project on correspondence networks, this means that a researcher would define three types of objects in nodegoat: 'letter', 'person', 'city'. Each object relates to an other object via relations (e.g. a letter relates to persons to identify the sender/receiver and this letters has been sent from/received in a city). In an extended research process, researchers could also define themselves as objects in the dataset, their sources or other datasets.[2] Due to the focus on relations and associations between heterogeneous types of objects, the platform is equipped to perform analyses spanning multitudes of objects. By enriching objects with chronological and geospatial attributed associations, the establishment and the evolution of networks of objects is inherently contextualised. In nodegoat, these contexts and sets of networked data can be instantly visualised through space and time.

This open-ended approach makes nodegoat different from tools like the Social Networks and Archival Context Project[3], Alan Liu's Research Oriented Social Environment[4], the Software Environment for the Advancement of Scholarly Research[5], Prosop[6], or tools with a main focus on coding of qualitative data as seen in various computer-assisted qualitative data analysis software. With its object-oriented approach, nodegoat facilitates the aggregation of collections, coding of texts, and analysis of networks, but models these methods towards the creation and contextualisation of single objects that move through time and space.

The analyses performed by nodegoat and the visualisations produced in nodegoat allow scholars in a variety of disciplines within the humanities to explore new research practices and methodologies. Joep Leerssen of the University of Amsterdam uses nodegoat for his project 'SpInTime – Dynamically visualizing how cultural patterns, networks and exchanges evolve in space and time'. By using nodegoat's data management and visualisation functionalities, SpInTime 'aims to map the dissemination of cultural nationalism across Europe by charting cultural patterns and networks as they evolve over time'.[7] The Ghent Center for Digital Humanities uses nodegoat to map conference attendance in the long nineteenth century.[8] In June 2014, students of UNIKA university in Semarang Indonesia used nodegoat during a workshop organised by NIOD. During this workshop, they interviewed survivors of anti-communist violence and built interactive mappings of an infrastructure of violence in nodegoat.[9]

In our presentation, we will introduce nodegoat's functionalities and explore two exemplary projects. Next, we will present a workflow in nodegoat to show how humanities researchers can make use of nodegoat for their own research questions.

---

1 http://nodegoat.net/about, http://historicalnetworkresearch.org/?topic=nodegoat-faq

2 P. van Bree, G. Kessels, 'Mapping Memory Landscapes in Nodegoat', in: L. Aiello, D. McFarland, Social Informatics. Lecture Notes in Computer Science (Heidelberg, 2015) 274–78. http://dx.doi.org/10.1007/978-3-319-15168-7_34

3 http://socialarchive.iath.virginia.edu/snac/search

4 http://liu.english.ucsb.edu/rose-research-oriented-social-environment/

5 http://www.seasr.org/

6 http://www.prosop.org/

7 http://spinnet.eu/spintimemappings and http://romanticnationalism.net/

8 http://www.tic.ugent.be/?q=VRE_description

9 http://www.niod.nl/en/projects/memory-landscapes-and-regime-change-1965-66-semarang

Assessing lexical quality of a digitized historical Finnish newspaper collection with modern language technology tools

Kimmo Kettunen and Tuula Pääkkönen
National Library of Finland, Centre for Digitization and Conservation
kimmo.kettunen@helsinki.fi, tuula.paakkonen@helsinki.fi

Digitization by means of scanning and optical character recognition (OCR) of both hand-written and printed historical material during the last 10–15 years has been an ongoing academic and non-academic industry. Most probably this activity will only increase in the ongoing Digital Humanities era. As a result of past and current work we have lots of digital historical document collections available and will have more of them in the future.

The National Library of Finland has digitized a large proportion of the historical newspapers published in Finland between 1771 and 1910 (Bremer-Laamanen 2014; Kettunen et al. 2014). This collection contains approximately 1.95 million pages in Finnish and Swedish. Finnish part of the collection consists of about 2.39 billion words. The National Library's Digital Collections are offered via the *digi.kansalliskirjasto.fi* web service, also known as *Digi*. Part of the newspaper material (years 1771–1874) is also available freely downloadable in The Language Bank of Finland provided by the FinCLARIN consortium[1]. The collection can also be accessed through the Korp[2] environment that has been developed by Språkbanken at the University of Gothenburg and extended by FIN-CLARIN team at the University of Helsinki to provide concordances of text resources. A Cranfield style information retrieval test collection has also been produced out of a small part of the Digi newspaper material at the University of Tampere (Järvelin et al. 2015).
The web service *digi.kansalliskirjasto.fi* contains different material besides newspapers, including journals, and ephemera (different small prints). Recently a new service was created: it enables marking of clips and storing of them to a personal scrapbook. The web service is used, for example, by genealogists, heritage societies, researchers, and history enthusiast laymen. There is also an increasing desire to offer the material more widely for educational use. In 2014 the service had over 10 million page loads. User statistics show that about 88.5 % of the usage of the Digi comes from Finland, but a 11.5 % share of use is coming outside of Finland.

Quality of OCRed collections is an important topic in digital humanities, as it affects general usability and searchability of collections (Holley, 2008, Tanner et al., 2009). There is no single available method to assess quality of large collections, but different methods can be used to approximate quality. This paper discusses different corpus analysis style methods to approximate overall lexical quality of the Finnish part of the Digi collection. Methods include usage of parallel samples and word error rates, usage of morphological analysers, frequency analysis of words and comparisons to comparable edited lexical data. Our aim in the quality analysis is twofold: firstly to analyse the present state of the lexical data and secondly, to establish a set of assessment methods that build up a compact procedure for overall quality assessment after e.g. re-OCRing or post-correction of the material. In the discussion part of the paper we shall synthesise results of our different analyses.

Our results show, that about 69 % of all the word tokens of the Digi can be recognized with a

1https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/KielipankkiAineistotDigilibPub
2https://korp.csc.fi/

modern Finnish morphological analyser. If orthographical variation of *v/w* in the 19th century Finnish is taken into account and number of out-of-vocabulary words (OOVs) is estimated, the recognition rate increases to 74–75 %. The rest, about 625 M words, is estimated to consist mostly of OCR errors, at least half of them being hard ones. 1 M most frequent word types in the data make 2.043 billion tokens, out of which 79.1 % can be recognized. If words that occur only once in the data (hapax legomena) are analysed, 98 % of them are unrecognized by morphological software.

The lexical quality approximation process we have set up is relatively straightforward and does not need complicated tools. It is based on frequency calculations and usage of off-the-shelf modern Finnish morphological analyzers. Even though we have done the estimation now in a partially automatized way, it is possible to automatize the operation completely. It is also apparent that we need to be cautious in conclusions, as different data are of different sizes which may cause errors in estimations (Baayen 2001; Kilgariff 2001). However, we believe that our analyses have shed considerable light into quality of the Digi collection.

Acknowledgements

This work is supported by European Regional Development Fund (ERDF), Leverage from the EU, 2014–2020.

References

Baayen, H. (2001). *Word Frequency Distributions*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Bremer-Laamanen, Maj-Lis, 2014. In the spotlight for crowdsourcing. Scandinavian Librarian Quarterly 1: 18–21.

Holley, Rose. 2008. How good can it get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. D-Lib Magazine March/April 2009. http://www.dlib.org/dlib/march09/holley/03holley.html

Järvelin, Anni, Keskustalo, Heikki, Sormunen, Eero, Saastamoinen, Miamaria and Kettunen, Kimmo, 2015. Information retrieval from historical newspaper collections in highly inflectional languages: A query expansion approach. *Journal of the Association for Information Science and Technology*. http://onlinelibrary.wiley.com/doi/10.1002/asi.23379/abstract

Kilgariff, A. (2001). Comparing Corpora. *International Journal of Corpus Linguistics,* 6(1), 97–133.

Kettunen, Kimmo, Honkela, Timo, Lindén, Krister, Kauppinen, Pekka, Pääkkönen , Tuula and Kervinen, Jukka, 2014. Analyzing and Improving the Quality of a Historical News Collection using Language Technology and Statistical Machine Learning Methods. IFLA World Library and Information Congress, Lyon. http://www.ifla.org/files/assets/newspapers/Geneva_2014/s6-honkela-en.pdf

Tanner, Simon, Muñoz, Trevor and Ros, Pich Hemy, 2009. Measuring Mass Text Digitization Quality and Usefulness. Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive. D-Lib Magazine July/August http://www.dlib.org/dlib/july09/munoz/07munoz.html

Unlocking a Finnish Social Media - In Search of Citizen Mindscapes

Krista Lagus, Minna Ruckenstein, Mika Pantzar, Jussi Pakkasvirta, Marjoriikka Yli-Siurua
University of Helsinki, Finland
krista.lagus@helsinki.fi, jussi.pakkasvirta@helsinki.fi

Social media services have become a crucial part of everyday conversational landscape in the past fifteen years. The combination of social scientists research questions, language technology and advanced data analysis methodologies offer thought-provoking possibilities for the study of social media sites.

Suomi24 is Finland's largest social media and it is also one of the biggest non-English online discussion forums in the world. Every month, 86% of Finns using the Internet visit the site (source: TNS metric) – at least to find out an answer to a question they have googled. Due to anonymity the discussions are topic-oriented rather than openly promoting the identity formation efforts of the author. The material opens unique perspectives for studying the qualitative and quantitative aspects of conversations and the dynamics of interaction, be it emotionally supportive, everyday problem-solving, or conflict-oriented.

We introduce the Citizen Mindscapes initiative, a research collective for social media analysis focusing on Suomi24, which recently obtained funding from the Digital Humanities programme of the Finnish Academy of Sciences. The research initiative brings together scholars from social sciences, digital culture, welfare sociology, language technology, and statistical data analysis. The aim is to build long-term interdisciplinary collaboration that develops into new ways of exploring social and political issues and conversational dynamics.

Citizen Mindscapes collaboration tackles the Suomi24 forum from three complementary perspectives. Firstly, it examines the digital culture of reading and writing, including the social shaping of users of social media. Secondly, the research initiative develops visual tools and statistical data analysis methods essential for analyzing and researching discussion forums, and thirdly, a few exemplary research questions are studied more thoroughly, such as the different types of micro interaction found in the forum, for instance, how heated debates might turn into political actions, or how to detect emotional waves from the discussions. The project utilizes methodologies and research insights from various research traditions for the study of a shared object, the Suomi24 forum data.

The initiative systematically applies collaborative processes for promoting open science, open data and open source. As part of the project, we facilitate the production of data sets tagged for various purposes. Our intent is that the data sets and the research tools developed are integrated in the Language Bank of Finland in collaboration with Fin-Clarin infrastructure and CSC Computing Centre. This will open up new research possibilities for a much larger research community.

The presentation describes some of the first steps and findings of the Citizen Mindscapes research frontier. Together with a network of collaborators, including Fin-Clarin researchers, CSC, Methods Centre of University of Helsinki, and the data owner Aller, we have opened the Suomi24 chat forum data, covering time period from the beginning of 2001 to May 14, 2015. The data are now available through the Language Bank and downloadable in total in JSON format. Thanks to Fin-CLARIN, searches on the data set (at first a subset, and later the full data) can be performed using Korp. Access to the data is managed by the Language Bank, and granted for research use.

Overall, the JSON data set contains about 53 million comments in 6,8 million threads, containing altogether 2,2 billion words (after removing duplicate entries). The data set divides into 21 sub-forums and several levels of sub-sub-forums. The three most active sub-forums are 'Society' with 26% of comments and 30% of words, 'Relationships' with 13% of comments and 10% of

words, and 'Local communities', 9% of comments and 7% of words.

Ongoing efforts, exploring research questions as collaborations between social science, humanities, data analysis, and programming, will be discussed as examples of dynamic scholarly processes and shared aims. The examples include the examination of everyday and weekly rhythms detected in the social media writings; the trends of emotional word usage over time in different discussion forums; and specific topics of interest, such as people's discussions around money, food, and health. It is our hope that reporting such pilot case studies facilitate the development of research and visualization tools that may be useful for the research perspectives of humanities and social scientists, thereby closing the gap between digital tool developers and scholars interested in everyday conversational landscapes.

References:

Krista Lagus, Minna Ruckenstein, Mika Pantzar (2015). Keskustelun tunneaallot – Suomi24 -avoimen datan hanke**.** *Tieteessä tapahtuu*, 33:6, 2015. http://ojs.tsv.fi/index.php/tt/article/view/53327

Suomi24 data set (2015). Suomi24-2015-05-25_JSON, Language Bank. https://korp.csc.fi/download/Suomi24/

Icelandic Saga Map Project: Digitally Mapping Medieval Icelandic Literature

Emily Lethbridge
University of Iceland
emily@hi.is

Icelandic Saga Map (ISM) (http://sagamap.hi.is) is a new DH tool and resource which uses GIS technology to link texts of the medieval Icelandic sagas (*Íslendingasögur*) to an online map interface. Development of the ISM project has been led by Emily Lethbridge (together with Trausti Dagsson, Hjördís Erna Sigurðardóttir, Gísli Pálsson and Logi Ragnarsson), funded by Rannís (the Icelandic Research Council) and supported by the Stofnun Árna Magnússonar í íslenskum fræðum, Reykjavík, and Miðaldastofa (Centre for Medieval Studies), University of Iceland. The resource is online and open to all. All places named in the sagas are geo-referenced and hyperlinked to their location on a base-map. One or more sagas can be loaded onto the map so that the distribution and overlap of places in the sagas is made visible. The website and interactive map thus enables users to 'read' the sagas out of the landscape and to explore, visually, the relationship and interplay between the sagas and the Icelandic landscape, the landscape being, effectively, the sagas' dramatic stage.

The ISM project was recently awarded First Prize by the University of Iceland in its annual open "applied science" competition (Hagnýtingarverðlaun Háskóla Íslands) for its potential for utilisation within and beyond academia – in teaching and research, tourism, and the presentation of cultural heritage in various formats and contexts.[1] The utilisation of the resource by several different user groups at once highlights and demonstrates the flexibility of digital maps as DH tools. Such flexibility might be considered to be one of the hallmarks of GIS-based DH projects, and one of the most exciting things they have to offer. In addition to the ISM digital tool's flexibility, there is also

1http://english.hi.is/frettir/interactive_icelandic_saga_map_won_applied_science_prize

great potential for the linking of the ISM database to other databases with a spatial component that are concerned with Icelandic history, literature and material culture. Building infrastruture to make this kind of multi-disciplinary collaboration possible is one of the objectives of the CyberNABO working group (http://www.cybernabo.org/), of which ISM is part.

Future plans for ISM development include adding other genres of medieval Icelandic texts (*biskupasögur*/Bishops' sagas, *konungasögur*/Kings' sagas, *þættir*/short tales of Icelanders, and also *Landnámabók*/The Book of Settlements), as well as some of the 19th-century travel journals and desciptions of saga-sites that 'saga pilgrims' such as William Morris produced. Work has also begun on developing the interactive functions of the map, thus enabling users to query the data in different ways.

In the proposed paper, after a brief description and demonstration of the ISM tool, discussion will turn to some of the methodological and theoretical issues and challenges that have come to the fore over the course of developing ISM to date. One aim of the ISM project is to consider, in an Icelandic context, such challenges as are encountered by all those who attempt to map literature in which 'real' space and places coalesce and overlap with literary/imaginary space and places. Iceland and the Icelandic saga corpus is an extremely rich body of texts to consider in this respect. Not least, this is because of the nature of the reciprocal relationship between the landscapes and the stories rooted in them: this relationship, it could be argued, might be said to characterise, definitively, the sagas' preservation and transmission from the medieval period onwards. Using ISM as a case study, then, focus will be directed at some ways in which digital maps can be used as tools to approach traditional research questions from new perspectives, and to formulate new research questions pertaining to the transmission of literature in landscape contexts, and the intersection of real-and-imaginary worlds.

Mapping life-stories of the Moravian Church 1730-2000

Sverker Lundin and Trausti Dagson
Centre for digital humanities, University of Gothenburg
sverker.lundin@gmail.com, traustid@gmail.com

The Moravian Church was founded by Count Nicholaus Zinzendorf in 1727. A central idea of this new ”church” was that its members were encouraged to write an autobiography documenting their life story, a *Lebenslauf*. Its primary purpose was to articulate and preserve the ”inner journey”, the person's path from a state of sleeping unawareness to his or her rebirth in Christ. Besides being read at the funeral, these *Lebensläufe* were disseminated to parishes around the world, where they were read as a means of edification. Some of them became very popular and were printed. Within the church it therefore emerged what for its time was a relatively ”modern” system of print and translation. The *Lebensläufe* played a central role for the church as they portrayed the ”inner journey” of its members, which in turn constituted the very core of the ”inner church”. For this reason an archive was established in Herrnhut in 1762. The main purpose of this archive was to preserve the *Lebensläufe*. When the missionary in the New world had taken off in the latter part of the 18th century, a second archive was established in Bethlehem, Pennsylvania. Together, the archives in Herrnhut and Betlehem contain about 20 000 *Lebensläufe*, spanning the time period from the establishment of the archive in 1750 to the present day.

The proposed paper is a presentation of a project constituting the first step in a planned process of

digitalization, transcription, presentation and analysis of the *Lebensläufe* of the Moravian Church. Only a small fraction of the material is printed, and even less is currently digitized. We have therefore as a first step focused on the metadata that sits in digital form in the archives in Herrnhut and in Betlehem. These metadata contain *place* and *date*, of *birth* and *death*, information about *gender* and sometimes other information concerning the *Lebensläufe*.

We will present a web-portal focused on these metadata. In the centre of the portal stands an interactive map of the world. On this map, the user can visualize subsets of the material based on the information in the metadata. For instance it is possible to show: the birth places of all persons living in a certain period; the death places of all women (or men) coming from a certain place or area or; the relationship between birth place and place of death, shown as arrows, for a subset of the material.

The web-portal is designed to serve several purposes. It will show-case the material and function as a nodal point for the international research community working with *Lebensläufe;* It creates new possibilities of exploration of the material, especially its spatial and temporal dimensions; It facilitates identification of subsets that can be made the subject of specialized research projects involving digitalization; It can function as an entry-point into the material sitting in Herrnhut and Bethlehem, bringing the old-fashioned possibility of actually going there to look at the *Lebensläufe* to the attention of researchers.

The project has been mainly data-driven with the primary purpose of opening the material up for research. A host of ideas for what to do with the material has been constantly present however. One specific focus is on the Moravian church itself, its spatial distribution as well as its demography in terms of ethnicity, class and gender. The material indicates that the church included a wide variety of people, from slaves and artisans to nobles. Another use of the material is as a more general point of entry to life-worlds of Europe. Because of its heterogeneity, the material provides a wide range of perspectives. Many of the Moravians travelled, which makes the material into a rich source of information regarding migration, between different parts of the world, at different points in time. It has been suggested that the Moravians prefigured later developments, in psychology and education, concerning the self and its development. This, among other things, makes the material relevant for the history of ideas. On another note, the material opens up for various studies employing language technology. For instance, it would be possible to study the spatial distribution of sentiments.

An important question that we presently discuss how to design our database and our user interface to best cater for this wide variety of possible research questions. Should the project focus on *documents*, or on *persons*? Should it only include text, or also include pictures, such as portraits or paintings of buildings? Should only *Lebensläufe* be included in the database, or should it also include related materials, such as contemporary printed sources? To what extent should we try to to link the information about persons found in the archive to biographic lexica?

In our presentation we will explain the design choices we have made, but also invite to discussion concerning the possible futures of the project.

Digital Palaeography and the Old Swedish Script. Some Preliminary Results and Future Prospects

Lasse Mårtensson
University of Gävle
laeman@hig.se

Palaeography incorporates the study of old script. This discipline has a very long history, and it has often been seen as offering the most trustworthy criteria for identifying the scribe of a manuscript and for determining when a manuscript was produced (i.e. scribal attribution and dating). In the last years, digital methods have been used for investigating the script, and this field is often referred to as digital palaeography. The digital methods mostly have their origin in image analysis, and using these we can identify and measure script features that are not visible to the human eye.

In this paper, an investigation within the field of digital palaeography will be described where the focus is on the chronological change of the script. The purpose of this is to find criteria for dating manuscripts, of the most central areas of palaeography. The material being investigated consists of medieval charters in Old Swedish and Latin (the latter also produced in Sweden). In these, the script is to be classified as Gothic, a script style widely used throughout Europe during the High and Late Middle Ages. The point of using charters in an investigation searching for change over time is that the charters are self dated, thus representing firm points of departure for the chronological development. Once the features have been extracted, these can be applied to the manuscripts (medieval books), which in the normal case not are dated. This has not, however, been done within this investigation.

As stated, this is an attempt at extracting features of the script that is relevant from a chronological perspective, for instance for dating purposes. The material consists of ca. 11000 Old Swedish charters. The basic principle for this method is that the computer unsupervised looks for regularities in the script, i.e. without direct instructions from a human. The method is executed in four steps. First, the contours of the script are measured. Actually, the contours are the only factor that the features are extracted from, and it is in each of the pixels in the contours that the feature extraction is performed. In the second step, features are extracted, unsupervised, from the contours of the script, and the number of the extracted features is 300. The unsupervised approach means that the computer extracts the features without human instructions, and it is allowed to search for regularities on its own, so to speak. The disadvantage of such an approach is that we do not inform the computer of certain chronological features of the script that we actually know about, and these are therefore not accounted for in the investigation. But the great advantage of this approach is that the computer may identify a completely new set of script features that display chronological change, hitherto undetected by the research. In the third step, the extracted features are related to the year of production on a selection of the corpus (ca. 5 %). The computer thus trains on this selection of material where the year of production is given to the computer, and the computer relates the extracted features to the year of production of the charters. In the last step, the computer uses the extracted information on material where the year of production is not given to the computer, and an estimation of year is given. As the charters are dated, the estimation of the computer can be checked.

The datings turned out fairly well, even though the time span given was quite rough, plus + minus ca. 20 years. Thus, the computer has discovered some features that may be relevant from a

chronological perspective, and that are not related to the style as such, for instance the change from the older to the younger cursive (which has been in focus of previous datings on palaeographic grounds). A full picture of which these features are cannot, however, be given at the present stage of research, but a few preliminary results will be shown.

Emergent Environments

Anna Aleksandrovna Misharina, Roger Mähler, Johan von Boer
Umeå University
anna.misharina@umu.se, roger.mahler@umu.se

There is a new type of spaces that is becoming increasingly common within academia, especially in the Digital Humanities context. The physical spaces of this type are clearly distinct from traditional academic spaces such as a seminar room, a lecture hall, a laboratory and others. They are characterized by extensive use of technology, their multifunctional setup and their multidisciplinary agenda. One can find such spaces in Scandinavia, Central Europe, North America, Australia and other places around the world. Due to extensive use of technology, large investments are in place, for example, in 2012, more than five millions euro were dedicated for the development of such environments at Umeå University[1]. These spaces are envisioned to produce *emergent* outcomes, and hence, such spaces are named *emergent environments* here on (abbreviated EE). *Emergence* is a process of complex patterns or regularities arising from interactions among smaller or simpler entities that are unable to produce such phenomenon on their own[2].

Within the academic domain, knowledge production is the predominant expected emergent outcome. It is in many cases uncertain whether these spaces actually meet the initial purposes and, indeed, produce emergent outcomes. Temple (2007) states that more resources should be put into the evaluation of the new academic spaces. He also argues that currently there is no complete methodology, developed and applied, to analyze these spaces. There are theories and related methodologies that take into account how the *sociality* and the *spatiality* influences each other(e.g. Space-Syntax[3] developed by Bill Hillier) without focusing on the specifics of the academic spaces. Knowledge production is difficult to assess due to its immaterial nature. The number of publications is often used as a measure of knowledge production. It is, however, in most cases impossible to single out the point in time when the initial and fundamental ideas for the publication appeared, and the relation between the publication and the space is hence uncertain. The proposed approach acknowledges the complexities of the knowledge production process, and instead focuses on the material space and its ability to afford activities that lead to knowledge production. Obstfeld (2005) argues that connecting disconnected individuals and enhancing existing connections, which are influenced by physical space, are strong factors for producing knowledge. According to Allen (2007), the spatial layout can promote informal communication and chance encounters that are critical for creativity and innovation.

Emergent environments are hard to analyze in entirety because of their complex and fluid nature.

1More information can be found here: http://www.umu.se/nyhet//.cid245160

2Similar concept is defined by J. Bennett as *assemblages* (Bennett, 2010).

3There are similarities between one part of the current methodology and Space-Syntax developed by Bill Hillier (Hillier & Hanson, 1989; Hillier, 2007). However, this work disagrees with social-spatial dualism stated in Hillier's works, arguing for the post-humanist approach.

However, they can be modelled and analyzed when decomposed into smaller *situations*[4]. A situation is a *bounded context* (a notion borrowed from software architecture) defined by its cohesion: spatial cohesion (architectural and design elements affording same or common activity), temporal cohesion (a single activity at a time) and social cohesion (individuals are linked to one another within the group and to the group as a whole). It is crucial to define both a situation and its relation to other situations within the EE.

The proposed methodology allows us to create two models of the analyzed space: a model of intended use and of actual use. The two models are analyzed and compared to show how the current space corresponds to the intended purpose, and how it can be improved to gain added values. The methodology studies the original idea, intention and expected social values of the space. Other parts of the methodology are static analysis of the architectural plan, as well as observations and tracking of the environment. The comparative analysis of the models indicates if activities that take place in the environment conform, contradict or are indifferent to the programs implied in the space. A discrepancy between the models reveals the potential for improvements in the physical space, and that consequently can influence activities and give added values. Through one of the case studies, a brief overview of the entire methodology - with a more detailed description of the tracking system (partially) developed within this project - and its employment in the comparative analysis are presented.

References

Allen, T. J. (2007). Architecture and communication among product development engineers. *Proceedings of the 2007 IEEE Engineering Management Society. EMS - 2000 (Cat. No.00CH37139)*, *49*(2), 23–42. http://doi.org/10.1109/EMS.2000.872493

Bennett, J. (2010). Vibrant Matter: A Political Ecology of Things. Durham: Duke University Press. Retrieved from https://www.dukeupress.edu/Vibrant-Matter/index-viewby=subject&categoryid=28&sort=author.html

Hillier, B. (2007). *Space is the machine: a configurational theory of architecture*. Space Syntax. Retrieved from http://discovery.ucl.ac.uk/3881/

Hillier, B., & Hanson, J. (1989). *The Social Logic of Space* (Reprint). New York: Cambridge University Press. Retrieved from http://www.cambridge.org/catalogue/catalogue.asp?isbn=0511839472

Obstfeld, D. (2005). Social Networks, the Tertius Iungens Orientation, and Involvement in Innovation. *Administrative Science Quarterly*, *50*(1), 100–130. http://doi.org/10.2189/asqu.2005.50.1.100

Temple, P. (2007). Learning spaces for the 21st century: a review of the literature. Retrieved from http://www.heacademy.ac.uk/projects/detail/litreview/lr_2007_temple

Thiel, P. (1996). *People, Paths, and Purposes: Notations for a Participatory Envirotecture*. Seattle: University of Washington Press. Retrieved from http://www.amazon.com/People-Paths-Purposes-Participatory-Envirotecture/dp/0295975210

4Similar concept of dividing space into smaller entities, *subspaces*, is found in works of Philip Thiel (Thiel, 1996). However, his works were directed towards notating spaces in order to understand their spatial, material and phenomenological qualities, whether this work aims to relate more layers with physical space as one of them.

Cross border access for researchers to source material by ECL

Christopher Natzén
National Library of Sweden
christopher.natzen@kb.se

The paper will discuss questions related to a pilot that the National Library of Sweden (KB) set up in order to investigate access to its collections, concerning both audiovisual and text based content. The aim is to, with support of existing legislation and within the limits of the current budget, create an infrastructure and business model for researchers that enables cross border access to the library's collections. This includes material that is under copyright.

Web-based access to source material deepens and simplifies the work for several research fields. Many infrastructural projects are run today with a specific scientific question in mind. This work distinguishes itself in so far as it does not focus on any specific question but how to give access to source material regardless of field of research.

For this purpose KB launched a national pilot in 2015 of giving researchers at two universities, Lund University and Umeå University, streamed access to the library's collections. This was followed by a letter of intent between KB, Copyswede and Kopiosto. It was signed on 5 November 2015 and opens up the door for the world's first cross border access agreement under an Extended Collective License (ECL). This will give researchers at Åbo Akademi access to KB's collections under the pilot scheme while in the future researchers at Lund and Umeå universities will have access to Åbo Akademi's collections.

The intention is to create a global cross border service of making available source material to the collections of an archive or a library over the Internet with streaming method by ECL. The making available of source material with a streaming method is done in manners and formats designed to not allow permanent storage or copying by the recipient or access by unauthorized persons. However, anyone with correct authorization may have access regardless of where they currently are located. This should also work cross border and the location should not impact negatively on access. Technology makes for global access, national borders is of little relevance for researchers and an infrastructure need to acknowledge these facts. The necessary technical costs should be kept low in order to allow archives and libraries, with lower possibilities to develop new technology, to be part of the service. The planned service is therefore demand led using existing technology allowing for archives and libraries to give access to its collections without the need to digitize beforehand.

As the infrastructure in theory allows for access to an archive and library's entire collections there will be materials that fall under the Data protection regulation. An infrastructure will have to be tuned to allow access and at the same time safeguard the interest of the general public whose personal, and in many cases sensitive information, is to be found in archives and libraries. The paper will show the work done so far in order to discuss issues that a project like this need to address, such as copyright and data protection legislation as well as technical issues like authentication and workflows.

Examples will be taken from KB's work of giving streamed access to both its printed material and audiovisual content. From a Swedish national perspective the legal basis is relatively

straightforward. Since November 2013 KB can negotiate and enter into a collective licence which is by law extended to non-members of the Collective Management Organizations (CMO) – thus referred to as an extended collective licence (ECL). An ECL allows KB to digitize its collections and provide internet access to it even though the collection contains works of unknown right holders (Orphan Works) as well as works of right holders who are not members of the CMO. However, how this is to be implemented at a global level, even at a EU level, is more complicated. When content is made available from a cultural heritage institution across borders diverging legislation collides. In some countries, such as in the Nordic nations, there are laws allowing extended collective licenses nationally making agreements fairly easy to reach. But, and as the pilot work so far has shown, as more countries are included even describing prof-of-concepts are difficult.

From a technical perspective, there are also many questions that need to be addressed in a possible solution. The paper will draw examples from issues in the pilot both internally at KB and externally, including but not limited to; solve the necessity of having direct access to KB's own network; create faster availability - shorter time span from ordered materials to access; increased availability to access material externally means that content should be viewable through different types of interfaces (computers, phones, tablets, etc.); increased user base puts demand on server capacity and performance; increased safety and opportunities to put restrictions based on the license; greater need for statistics and monitoring (e.g., for quality reasons and to prevent misuse of the service).

http://www.mynewsdesk.com/se/kungliga_biblioteket/pressreleases/kungl-biblioteket-och-aabo-akademi-inleder-samarbete-foer-digital-tillgaang-till-bibliotekens-samlingar-1215983

http://www.mynewsdesk.com/se/kungliga_biblioteket/pressreleases/kb-undertecknar-avsiktsfoerklaring-som-oeppnar-doerren-foer-vaerldens-foersta-avtal-om-graensoeverskridande-fjaerraccess-1247722?utm_campaign=send_list&utm_medium=email&utm_source=sendgrid

http://www.copyswede.se/2015/11/genombrott-i-att-tillgangliggora-tv-och-radioarkiv-for-forskning/

Uses and Extensions of Digital Resources for Multimodal Communication in an International Study Program at the University of Copenhagen

Costanza Navarretta
University of Copenhagen
costanza@hum.ku.dk

1. Introduction

This paper addresses the need for multimodal annotated resources for studying human-human and human-machine interaction and the re-use and extension of existing Nordic digital multimodal resources in an international master program, IT and Cognition, at the University of Copenhagen. Face-to-face communication is multimodal and includes at least two modalities: auditory (speech) and visual (gestures). Gestures comprise all types of body behavior such as head movements, facial expressions, gaze, body posture and hand gestures. The relation between speech and gesture is complex and depends on numerous factors such as the communicative situation, the culture, the

number ni maírtinof participants, the physical settings. Even though the number of video- and audio-recorded monologues and dialogs of different type grow every day on the internet, there are no agreed upon methodologies for analyzing and processing these data. The past two decades a number of Nordic and European projects have developed methodologies for the construction, annotation and use of multimodal corpora, but ethical and copyright issues make it difficult to share the corpora, especially if they consist of naturally occurring monologues and dialogues. Furthermore, manual annotation of multimodal corpora is time consuming and automatically annotated gestural information is coarse-grained and often not related to speech content. Therefore, there is a lack of sharable annotated resources for studying and modeling multimodal behavior to be used in human-computer interfaces and this is a problem for research and education.
In the following, I present shortly the study program addressed by this paper and the multimodal resources available to the students which have been developed in various Nordic and European projects. Then I describe how these resources have been used and extended by the students and discuss positive as well as problematic aspects of multimodal resources produced by students.

2 The education and the available resources

The IT and Cognition international master program is about modeling, developing and evaluating cognitive-based IT technologies and resources focusing on human language and vision. The cognitive aspects behind multimodality in human-human communication and their relevance for the implementation of multimodal human-machine interfaces are one of the subjects addressed in the program. Available annotated multimodal data of both human-human and human-machine interaction are few and have been mostly produced the past decade.

Nordic and European projects which have addressed the construction and annotation of multimodal corpora comprise the Danish Clarin and VKK projects, the Nordic MUMIN and NOMCO projects, and parts of large European infrastructure and/or research and educational projects for digital humanities (Clarin, CLARA). In particular, the so-called MUMIN annotation framework for formally annotating multimodal corpora was proposed by the Nordic funded network, MUMIN. The framework proposes shape and function features describing communicative gestures and their relation to speech. According to the framework, and the schemes which implement it, the shape and function of gestures are independent features and gestures can have multiple functions. The MUMIN scheme has been refined, extended and applied in numerous projects. Examples of these extensions related to a Nordic language, Danish, are: a) the annotation of deictic gestures and their relation to anaphora in the Danish CLARIN multimodal corpus (Navarretta 2011), b) the annotation of hand gestures representing events in the CLARA project (Lis 2014), c) the annotation of emotions in the NOMCO first encounters corpora (Navarretta 2012).

These resources have not only been extensively used in research for analyzing, comparing and automatic predicting the function of gestures and their producers in different conversations and languages, but they have also been used in teaching and have even been extended in master theses at the IT and Cognition master program.

2. The creation of digital multimodal resources by students

The existing multimodal resources are used in teaching, but since they only focus on some aspects of multimodal communication, or address languages not understood by the international master students, the students are often obliged to create their own corpora and annotate specific features in the corpora.

The most extensive corpus collected by students consists of 12 first encounters between the NAO robot and cognitively disabled adults as well as 4 encounters between the robot and a control group

of adults with no disabilities. The first encounters were collected in cooperation with the municipality of Copenhagen by two master students, Eilersen (2013) and Lockert (2014). The two students also annotated smiles and feedback head movements in part of the corpus using an adapted version of the MUMIN scheme provided by their supervisor. This work has been presented in both international (Eilersen and Navarretta 2014) and national conferences (Eilersen 2014), but the corpus is only available to a restricted number of researchers due to ethical constraints.

Another multimodal corpus of 10 Danish narratives was collected by another student as part of her master thesis, Damkjær (2015), who adapted the CLARA scheme for analyzing event representing hand gestures and applied it to the Danish narratives in order to compare the work by the CLARA PhD student (Lis 2014) on the use of iconic hand gestures in in Polish and English narratives with the gestures in the Danish data. The Danish narratives can be used for research.

The extended MUMIN emotion scheme (Navarretta 2012) has been extended and applied in a master thesis investigating the flow of emotions in problem solving game (Christensen 2014). A multimodal corpus was collected in which the participants' behavior while they were playing the game was collected and partially annotated. The data comprised videos of the participants' facial expressions, game logs, galvanic skin response data, and questionnaire responses in which the participants described their feelings during the game

3 Discussion

The resources produced by Nordic and European infrastructure and research projects have been used and extended by master students at the University of Copenhagen. Furthermore, master thesis students have also collected new resources, and some of them will be made available for research and teaching. By creating their own data, the students learn in practice how to collect multimodal data aimed to investigate specific phenomena. They also learn to deal with confidentiality and ethical issues, and to annotate the chosen phenomena adapting or extending existing annotation frameworks. They also compare existing research resources with the newly produced resources which address other languages or other types of settings, e.g. human-human vs. human-robot interactions. However, even though some of the resources produced by students can be further used in research, the creation of multimodal resources from scratch by student is not unproblematic.

First of all, the process of creating a corpus is very time consuming and thus, in the end, the students do not have enough time to analyze in depth the phenomenon which motivated the data collection. Secondly, the participants who are recorded in the videos often pose more restrictions to the use of the data than they would do if the data had been collected in research projects because they are not sure about what will happen with the recordings and who will have access to them after the master thesis work has been completed. A third problematic aspect is the quality of the annotations since they are restricted to specific phenomena and are in most cases produced by one student and thus, they have not been checked and inter-coder agreement tests have not been applied to the annotations. Finally, in the cases in which the data are freely available for further research, they still need to be processed in order to be made available in common research infrastructures. Since the owners of the data in most cases get employed outside the university and miss interest in their master thesis data, it is necessary to have manpower to take care of the recordings and of ensuring the quality of the annotations.

References

P.A. Winther Christensen 2014. *Exploring Cognitive-Affective States with the Puzzle-Platform Game Antichamber*, Master thesis, IT and Cognition, University of Copenhagen.

C. Wejlgaard Damkjær 2015. *Hand Gesture Representation of Action Events in a Danish Corpus of Narratives*, Master thesis, IT and Cognition, University of Copenhagen.

T. J. Eilersen 2013. *A Multimodal Corpus of Communicative Behaviour of Cognitively Disabled Adults during HRI with an Anthropomorphic Robot*, Master thesis, IT and Cognition, University of Copenhagen.

T. J. Eilersen 2014. Communicative behaviors during human-robot interaction for cognitively disabled users. Presented at *Workshop on Embodiment in Multimodal Communication*, University of Copenhagen.

T. J. Eilersen and C. Navarretta 2014. A Multimodal Corpus of Communicative Behaviors of Disabled Individuals during HRI, *LREC Workshop Multimodal Corpora: Combining applied and basic research target*s, Reykjavik, Island, 27 May 2014, pp. 33-36. http://www.lrec-conf.org/proceedings/lrec2014/index.html

M. Lis 2014, *Multimodal representation of entities, a corpus-based investigation of co-speech hand gesture*, PhD Dissertation, University of Copenhagen.

M. Lis and C. Navarretta 2013. Classifying the form of iconic hand gestures from the linguistic categorization of co-occurring verbs. In: *Proceedings of the MMSym'13*, Malta, 2013, pp. 41-50.

M. Lockert 2014. *A Corpus of Smiles and Laughter, Nonverbal Phatic Behaviours in a Woz-Based HRI, A Comparison of Cognitively Disabled and Healthy Users*, Master thesis, IT and Cognition, University of Copenhagen.

C. Navarretta 2011. Anaphora and gestures in multimodal communication. In Hendrickx, Branco, Lalitha Devi and Mitkov (Eds.) Proceedings of the 8th DAARC 2011, Ediçoes Colibri, 2011 pp.171-181. http://www.cst.dk/costanza/papers/daarc_2011_navarretta.pdf

C. Navarretta 2012. Annotating and Analyzing Emotions in a Corpus of First Encounters. In *Proceedings of the 3rd IEEE CogInfoCom*, Kosice, Slovakia, 2012, pp. 433-438. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6422020

A cross cultural examination of the correlation between ‘selfies’, self-esteem and introversion in young people living in Ireland and Hong Kong

Sarah Stephanie Ní Mháirtín
University of Helsinki
sarah.ni-mhairtin@ucdconnect.ie

This study is the first of its kind to explore the modern phenomenon of the use of ‘selfies’ – photographs of oneself- in social media and its relationship with an individual’s personality

characteristics. Growing academic interest within the field of psychology, concerning private use of social media and the correlation between internet use and an individual's personality traits (e.g. Amichai-Hamburger, Wainapel, & Fox, 2002; Orr, et al., 2009) is becoming increasingly evident. Young adults living in the modern age are appropriately dubbed, the 'Net Generation' (Jones & Fox, 2009) and media-use is a central figure in their social, emotional and cognitive development (Roberts, Foehr, & Rideout, 2005). The present study aims to address the gap in 'selfie' research, a trend that is snowballing as social networking develops a more personalised approach, with greater emphasis on visual self-presentation (Hogan, 2010; Van Dijck, 2013).

This study examined the correlation between introversion, self-esteem and the intensity of the use of selfies, two traits consistently shown to be related to internet usage - particularly Social Networking Sites (SNS). Introversion was chosen as a focus of the study, following the recommendation of Ross et al. (2009) to examine narrow characteristics of personality when conducting internet related research. As with most research concerning personality, cross cultural differences are highly likely (e.g. Paulhus, Duncan & Yik, 2002; Schmitt, Allik, McCrae, & Benet-Martínez, 2007), thus prompting the inclusion of the Hong Kong population. This correlation was also examined in both Hong Kong and Irish populations, given the lack of cross-cultural research in the area of personality and internet use. This focus was selected following observations in the literature highlighting the greatest diversity between east-west cultures. Hong Kong in particular was chosen as a nation deemed to be in direct contrast to Ireland. Compared to a 'Western' (and particularly an Irish population), Hong Kong as a Far East Asian society is typically perceived as more introverted (Furnham & Cheng, 1999; Schmitt et al., 2007) with lower self-esteem (Crocker, Major, & Steele, 1998; Spencer- Rodgers, Peng, Wang & Hou, 2004). Furthermore, its residents report heightened internet use and engagement with social networking sites (Alhabash, Park, Kononova, Chian, & Wise, 2012). Significant correlation between selfie intensity and the personality dimension of introversion and between selfie intensity and self-esteem were expected. Significant differences between Hong Kong and Irish samples in levels of introversion, self-esteem and selfie- intensity were also predicted.

Data were gathered using a convenience sample, contacted by means of SNS. Participants consisted of 186 young people, living in either Ireland or Hong Kong, matched for age and with a comparable mix of males and females (see Table 1 below for sample demographics). Of those living in Hong Kong, 45 identified as Hong Kong Chinese nationals, six as Mainland Chinese nationals, two as Irish and 33 as 'other' (see Table 2). In contrast, the entire sample from Ireland reported their nationality as Irish. Ethical approval for the study was granted by the Undergraduate Research Ethics Committee, School of Psychology (UREC-Psy). Participants were obliged to complete a consent form and confirm that they were over the age of 18. Participants were assured of the confidentiality of their responses and provided with avenues of care to follow should they experience distress during or after participation.

Table 1
Sample demographics for participants living in Ireland and Hong Kong (N= 186)

	Ireland	Hong Kong
N	99	87
Male (n)	39	23
Female (n)	60	64
Mean Age (*SD*)	21.89 (4.7)	21.14 (4.46)

Note: SD = Standard Deviation

Table 2
Nationality for participants living in Hong Kong

	Nationality
Chinese (Hong Kong)	45
Chinese (Mainland)	6
Irish	2
Other *	33

* Of these, 5 participants identified as Indian and 11 as other Asian nationals

Participants reported their levels of introversion and self-esteem via standardized scales: the McCroskey Introversion Scale (based on the work of Eysenck (1970; 1971) and designed to exclude measures of communication apprehension (McCroskey, 1984; Richmond & McCroskey, 1998)) and the Rosenberg Self-Esteem Scale (1965). Participants also provided information regarding their engagement with selfies and social media, via the Selfie Intensity Scale, a modified version of the Facebook Intensity Scale (Ellison, Steinfield & Lampe, 2007) - a quantitative measure of the degree of emotional attachment between an internet user and the SNS they engage in, along with the site's influence on an individual's everyday routine. Four basic questions relating to selfie use were also posed by the researcher. It should be noted that a pilot study was conducted prior to data collection to ensure that the Selfie Intensity Scale (adapted from the Facebook Intensity Scale) was valid and coherent. Twenty participants living in either Ireland or Hong Kong and matched for age and gender, participated and all feedback provided was used to clarify the final version of the Scale. Using two-tailed independent t-tests to ascertain differences between the two groups it was found that those living in Ireland and Hong Kong scored comparably with respect to introversion (t = -3.421; df = 181; $p > .05$),; self-esteem (t = -1.755; df = 180; $p > .05$), and; selfie intensity (t = -.426; df = 180; $p > .05$).Table 3 outlines the mean and standard deviation for all three measures.

Table 3
Scores for introversion, self-esteem and selfie intensity for participants living in Ireland and Hong Kong

	Ireland	Hong Kong
Introversion *Mean* (*SD*)	29.97(4.84)	32.44(4.93)
Self-Esteem *Mean* (*SD*)	10.13(3.15)	10.97(3.30)
Selfie-Intensity *Mean* (*SD*)	7.85(3.05)	8.04(2.90)

To explore the main hypotheses that selfie use is predicted by the personality characteristic of introversion and by self-esteem, correlational analyses were performed.

Figure 1 plots selfie intensity by introversion and shows a (negative) linear trend to the data, an assumption for Pearson correlations. Using a one-tailed Pearson correlation a significant negative relationship was observed between introversion and selfie intensity (r = -.218; $p < .05$). This represents a low correlation. From the r value, it is possible to calculate the measure of association (r^2 x 100), which indicates the percentage of variance in the data that is explained by the relationship between the two variables. In this example, the percentage of variance accounted for by the relationship between introversion and selfie intensity is 4.75%.

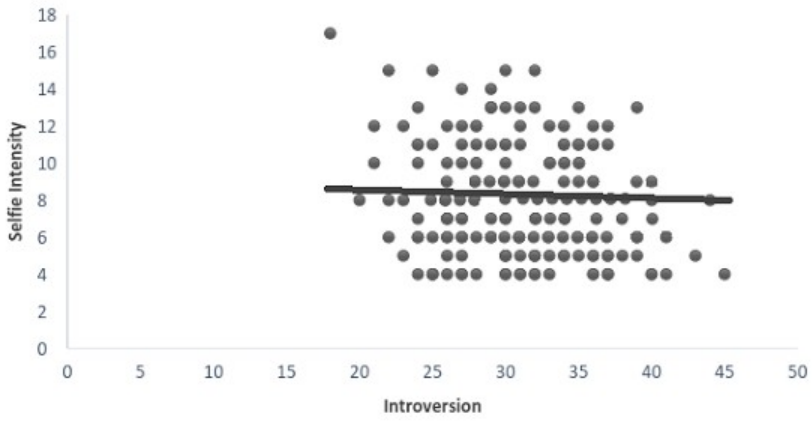


Figure 1. Correlation between selfie intensity and introversion

Figure 2 plots selfie intensity by self-esteem and displays a (positive) linear trend to the data, an assumption for Pearson correlations. Using a one-tailed Pearson correlation a significant relationship was not observed between self-esteem and selfie intensity ($r = .066$; $p > .05$).

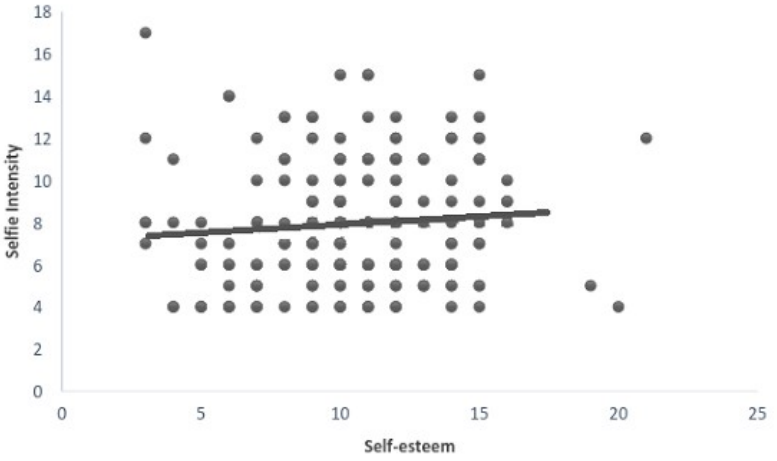


Figure 2. Correlation between selfie intensity and self-esteem

In addition to the analysis of selfie intensity, 4 basic questions posed by the researcher were examined in terms of their relationship with introversion. Each proved to be non-significant when correlated with introversion and self-esteem (see Table 4 for details).

Table 4
Correlation between 4 basic selfie questions, introversion and self-esteem

		Introversion	Self-esteem
Q5. How often do you use SNS?	Pearson Correlation	.021	-.094
Q6. Have you ever taken a selfie?	Pearson Correlation	-.044	-.036
Q7. How often do you post selfies?	Pearson Correlation	-.119	-.063
Q8. Who can view your selfies?	Pearson Correlation	.079	.003

Although previous literature pointed to a strong diversity between 'Eastern' and 'Western' societies, surprisingly, no significant differences were noted between the Hong Kong and Irish populations with regard to their scores for introversion, self-esteem and selfie intensity. The researcher points to the constant "Westernising" of Hong Kong (Chau, Woo, Gusmano & Rodwin, 2012; Wu & Ng, 2011) as a possible explanation of the noted similarity. As a result of this globalisation, Western preferred traits of extraversion (e.g. Grant, 2013; Hills & Argyle, 2001) and high self-esteem (Diener & Diener, 2009; Young & Morris, 2013) are often highly regarded and encouraged within a society that until now upheld traditions of introversion and reported low levels of self-esteem (Hong, Morris, Chiu & Benet-Martínez, 2000). In addition to this, Hong Kong is also becoming increasingly multi-cultural (Schnurr & Zayts, 2012), as seen by the mixed ethnicities of participants. This may also have influenced the noted change in mean personality levels, compared to previous studies. Studies also suggest that Hong Kong is on the brink of abandoning its former collectivist values. With economic growth, the society had developed a greater respect for individualistic traits and become more open to Western influences (Allen et al., 2007). It is highly likely that such increased levels of individualism have nurtured the formation of personalities similar to those typically found in Western cultures.

A weak correlation was reported between selfie intensity and introversion, but no relationship was found between selfie intensity and self-esteem. Thus, the correlation between individual characteristics and internet (and subsequently selfie) use may not be as significant in our rapidly developing modern society as recent research suggests. Other factors such as motivation and competence may be more relevant to selfie use than personality. Despite non-significant results, this study provides the first known attempt to further our understanding of the increasingly popular selfie and provides interesting, novel findings about youth culture.

Further investigations between various interactions with SNS and selfie-use/ intensity would have been beneficial, along with an examination of the motivation to take selfies (as has been done for SNS e.g. Ross et al., 2009). Further research could also examine the correlation between narcissism and selfie use, as the literature has noted that individuals who score highly on narcissism tend to engage in much self-promotion on their Facebook profiles (Buffardi & Campbell, 2008). An investigation into employers' attitudes towards selfie use and selfie posting, similar to research by Acquisti and Gross (2006) or Lampe, Ellison and Steinfield (2008), could be useful for young people entering the job market. Such information could be useful for companies who advertise on SNS, particularly on sites which encourage users to upload their selfies. This may assist such organisations in enhancing their advertising effects and tailoring their marketing campaigns to

match users' personalities- as recommended by Keng and Liu (2013). Furthermore, the results of this study show that the majority of participants have taken a selfie and uploaded it to a SNS. This in turn implies that individuals are becoming more absorbed with their appearance on SNS. This obsession may add to existing concerns regarding unhealthy levels of internet use and poor body image in young people (Tiggemann & Slater, 2013), in addition to self-objectification and the resulting clinical disorders (Slater & Tiggemann, 2011). The findings of the present study could be used to examine and address such issues in a practical setting.

Making web pages for TEI-P5 texts and connecting them with dictionaries

Jonas Nielsen
Society for Danish Language and Literature
jn@dsl.dk

Det Danske Sprog- og Litteraturselskab (DSL, Society for Danish Language and Literature) is an organization which has published historical texts and dictionaries in print since 1911. This talk will explain how we publish texts online and how we connect them with dictionaries.

We will focus on two sites containing medieval texts related to Denmark. The addresses of the web pages are middelaldertekster.dk and diplomatarium.dk. A substantial number of the texts on these sites are written in "Old Danish", others in Latin and Middle Low German.

The talk will be technical, but since demos will be shown, most people should be able to understand what it is about.

We imagine the two main parts of the talk will be as follows:

1) The medieval texts in question are marked up by our editors in an XML format called TEI-P5. We will talk about how we present texts in this format on a web page to people who normally read them in print. In addition to the presentation of the texts we will talk about searching and studying the original scans (facsimiles) of the document while reading the transcribed text. These features are already in production on the web pages mentioned.

If the listeners are interested, and time allows it, we will go into details regarding the database (eXist) and the web framework chosen (Pyramid — a python framework).

2) DSL has published most of their printed dictionaries on the site ordnet.dk, and added new ones not available on print. Our latest publication is an Old Danish->Danish dictionary. This dictionary (GO) is available at gammeldanskordbog.dk.

Simultaneously with the work on GO, we are working on a feature on middelaldertekster.dk which will look up each word from the text in the dictionary and show the result in a popup window (linking to a detailed page for the lemma).

It is a kind of "transparent" Google Translate. The user can read the text in the original language and look up the words he does not understand, without leaving the page he is looking at. You can go

to http://middelaldertekster.dk/dvaergekongen-laurin and click on a word to see an example.

If we have the time, we will go into detail about how the dictionary lookup popup is made. Each text is processed by a lemmatizer. The lemmatized text is then checked by a human editor using a specialized tool. The tool presents the editor for candidates with words in the dictionary, showing the sentence the words appear in, and the editor chooses the correct lemma. Both the tool and the lemmatizer are made in-house at DSL.

A description of the three projects in non technical terms goes as follows:

Diplomatarium Danicum (DD): The aim of this project is to publish medieval charters (diplomas) from Denmark as well as letters from outside Denmark that refers to Danish affairs. DD was founded in 1932 and since then more than 18,000 charters have been published covering the period between 789 and 1412. Newly published charters are available on diplomatarium.dk and not in print.

Studér middelalder på nettet (SMN): The aim of this project is to publish medieval texts that are not charters, and therefore normally longer than the texts from DD. The project is available at middelaldertekster.dk

Gammeldansk Ordbog (GO): The goal of this project is to publish an online Old Danish -> Danish dictionary. Work on the paper edition of this dictionary started in 1932. In 2010 we digitalized the work done so far on the web page gammeldanskseddelsamling.dk (Old Danish note collection). In January 2016 we made the dictionary available with the same interface as our other dictionaries on gammeldanskordbog.dk. The contents in the dictionary is, however, still incomplete.

Hur kan digitala samarbeten mellan forskare och minnesinstitutioner underlättas?

Reflektioner utifrån ett samarbete mellan historiker vid Lunds universitet och Universitetsbiblioteket i Lund.

Anna Nilsson Hammar
Lund University
anna.nilsson_hammar@hist.lu.se

Att presentera, synliggöra och tillgängliggöra samlingar på webben har blivit allt viktigare för den så kallade ABM-sektorn. Det kontinuerliga arbetet med att tillgängliggöra information om materialet i de befintliga bibliotekskatalogerna är givetvis en viktig del av detta arbete, men ökade krav ställs också på tillgång till digitaliserat material, överblickbarhet samt sök- och hittbarhet i digitala system. Forskare, inte minst inom de historiska vetenskaperna, har en viktig men inte alltid erkänd funktion att fylla i arbetet med digitala samlingar, visualisering och digitalisering av arkiv-, biblioteks- och museimaterial. Frågor kring urval och förmedling kräver inte sällan historievetenskaplig kompetens, så även berikning av metadata och kontextualisering av material. Digitala samarbeten mellan forskare och ABM-sektor kan få en rad positiva konsekvenser, bland annat vad gäller internationella forskningssamarbeten, allmänhetens tillgång till historia och forskares tillgång till och användning av material.

Detta paper tar sin utgångspunkt i ett samarbete mellan forskare vid historiska institutionen vid Lunds universitet och bibliotekspersonal vid Universitetsbiblioteket i Lund som resulterat i portalen *Guide till Universitetsbibliotekets specialsamlingar* (www.guide_specialsamlingar.ub.lu.se). Portalen är en statisk webblösning gjord i html och ett småskaligt svar på ett mer omfattande problem, nämligen att synliggöra vad som döljer sig i Universitetsbiblioteket samlingar, med särskilt fokus på material som i nuläget inte finns tillgängligt eller sökbart i befintliga databaser. Projektet pågick på deltid i sex månader och är ett exempel på hur mindre insatser kan få relativt omfattande resultat. Drygt 5000 sidor text har digitaliserats och finns tillgängliga (men inte sökbara) i portalen. Arbetet med att föra över dessa filer till den nya plattformen Alvin kommer att påbörjas inom kort. Projektet har också inneburit att mer resurser avsatts till digitalisering av handskriftsmaterial ur UB:s samlingar. Utifrån erfarenheterna av detta projekt vill jag diskutera *hur digitala samarbeten mellan forskare och minnesinstitutioner kan underlättas och förbättras.*

Fördelarna med interaktion och digitala samarbeten mellan forskare och minnesinstitutioner är många. Trots detta är ibland avståndet mellan forskarvärld och minnesinstitutioner onödigt långt. Jag vill i detta paper peka på ett antal faktorer som jag tror spelar en avgörande roll för möjligheten till digitala samarbeten och diskutera hur vi kan underlätta framtida satsningar.

- resursidentifiering
- behoven av kunskapsutbyten och kontinuerliga kontakter
- administrativ samordning
- digital infrastruktur vid respektive institution
- digital småskalighet och storskalighet

The Danish Thesaurus: semantic and thematic structure and its use in digital humanities

Sanni Nimb
Det Danske Sprog- og Litteraturselskab
sn@dsl.dk

A Danish thesaurus, *Den Danske Begrebsordbog* (Nimb et al. 2014) has recently been published by the Society for Danish Language and Literature (DSL). It presents 192,000 words and expression in 22 chapters and 888 thematic sections in which the words are presented in five word class subsections. The book is just one example of usage of the underlying, annotated XML manuscript where the vocabulary of the sections are grouped in subsections according to meaning, not word class, and furthermore formally annotated. In this talk I will describe the structure of the underlying XML manuscript and the types of semantic and thematic information which can be extracted from it. Furthermore, I will refer briefly to some experiments where this information has already proved its use, before I present some ideas on future exploitation of the thematic information in different projects at DSL.

The compilation of the thesaurus

The thesaurus was compiled on the basis of the 110,000 lemmas and expressions described in a corpus-based monolingual dictionary, *Den Danske Ordbog* (DDO, ordnet.dk/ddo). The hyponymy hierarchies established in *DanNet* of half of DDO's lemmas (Pedersen et al. 2009) were also exploited, making it easy to establish preliminary groups of words denoting e.g. 'furniture', 'clothes',

'animal' etc. in the relevant sections in the thesaurus. The three lexical resources are linked together at sense level, meaning that we are now able to combine the data in many different ways.

The thesaurus structure

The words and expressions in the 22 chapters and 888 sections of the XML manuscript are grouped in 8,300 subsections. For instance in the section “Nysgerrig” ('Curious') one subsection with words denoting the property 'to be curious', one with words for curious persons (*kødrand* ('circle of spectators'), *pilfinger* ('busybody')), and one with words for something making you curious (e.g. *globryllup* (wedding ceremony which you attend out of curiosity, as an uninvited guest), *teaser* ('teaser ad')). Each of the subsections contain formalized information in a header according to a restricted set of coarse-grained semantic types and semantic relations, established on the basis of our experiences with the compilation of the Danish WordNet (Pedersen et al. 2009). The formalized information makes it possible to identify groups of events, acts and properties as well as different hyponomy groups such as persons, artifacts, etc. throughout the XML manuscript. Fig. 1 describes the structure.
The structure also allows us to assign numbers representing thematic information at three levels to all words and expressions in the thesaurus, e.g. the noun *lilleputland* ('tiny country') belongs to one subsection of 8,300, one section of 888 (named “Stat, nation” ('State, nation')) and one chapter of 22 (named “Samfund” ('Society'). At the most fine-grained level the number is not yet connected to a topic in the form of a title, but we plan to identify one semi-automatically, based on the extraction of the first keyword of each subsection. We also plan to disambiguate the text strings in the 888 section titles, due to the fact that the strings in most cases occur as keywords with sense ids in the section, typically initiating one or two of the first subsections.

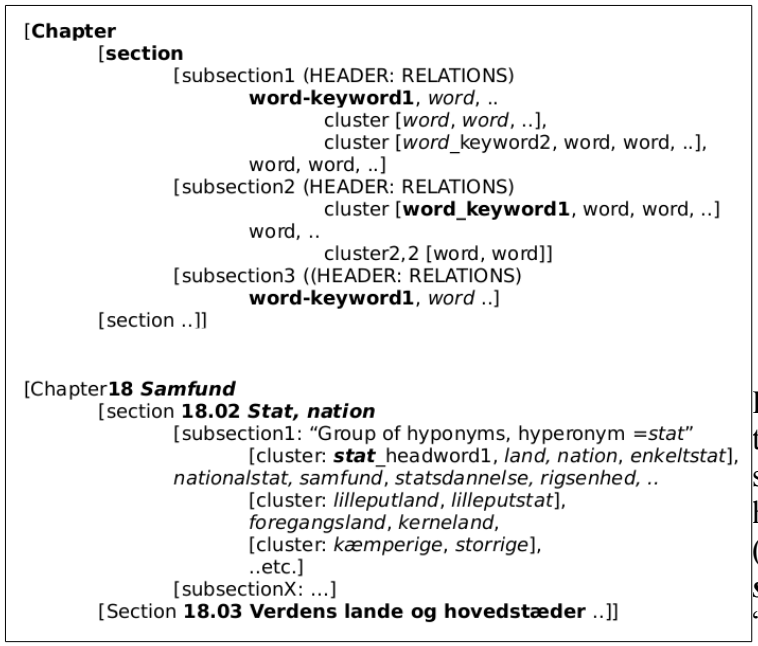


Fig.1: The XML structure of the thesaurus, exemplified with the subsection containing hyponyms of the noun *stat* ('state') initiated by the keyword ***stat***_headword1 in the section “18.02 Stat, nation” in the chapter “Samfund” ('Society').

Use of the XML structure and the annotations

The extractable semantic and thematic information has already been exploited in different projects. A Danish frame lexicon covering communication and cognition senses has been established on the basis of 'act' subsections (identified via their type number) of all sections covering these topics in the thesaurus, combined with the valency patterns of the identified vocabulary in DDO (Nimb &

Pedersen, forthcoming). The method proved to be quite efficient, and funding has been provided to extend the lexicon in 2016 and 2017. A method of supplying the Danish WordNet DanNet with more adjectives, as well as with information on the external argument of adjectives, based on information in the 'property' subsections in the thesaurus (also identified via their type number) is described in Nimb & Pedersen (2012). Experiments have also been carried out on the automatic transfer of the three thematic levels from the (at that time not yet completed) thesaurus data to DanNet. In the first test of the transferred data, information at the most coarse-grained level (the 22 chapters) was used to subdivide co-hyponyms in the WordNet thematically, and two percent of the assignments were manually judged in order to see to which extent the subdivisons of large hyponymi groups (eg. the many co-hyponyms of 'person') made sense. In spite of the coarse-grainedness, only 5 percent were not considered fully intuitive, and the experiment showed a considerable enrichment of the WordNet-data, see Nimb et al. (2013) for more details. Finally we should also mention that the printed thesaurus has been used to manually compare the lexicon of six medieval Danish romances of chivalry with two different provenances in order to obtain a clearer picture of the thematic differences between the texts (Akhøj Nielsen 2015).

Future plans

Apart from modern dictionaries, DSL also publishes dictionaries for elder Danish as well as classical Danish literature, in many cases on the basis of highly structured XML manuscripts. We consider the thesaurus a central lexical resource and plan to use it in several ways in the future. By creating (semi-automatic) links between the vocabulary of elder Danish dictionaries to the vocabulary of our modern Danish dictionary, DDO, the common ids between DDO and the thesaurus make it possible to transfer the thematic divisions to these dictionaries, opening up for new ways of using and presenting the elder vocabulary. Along the lines of the manual comparison of the words in medieval texts mentioned above, the idea is also to use the thesaurus to identify topics automatically in the digitized editions of elder Danish literature published by DSL (ADL.dk), allowing new types of search facilities in the texts. It implies of course a manual extension of the thesaurus with the elder Danish words that cannot be linked to a modern equivalent. As a first step towards our goals, we are involved in a research application with the University of Copenhagen aiming at identifying topics automatically in the digitized editions. In the project, results based on standard topic modelling methods will be compared with results where the thematic information from the thesaurus is added to the automatic process.

Literature

Akhøj Nielsen, Marita (2015): *Studies of vocabulary in lemmatized Old Danish texts* (speech at "Internationell konferens i östnordisk filologi", Dipartimento di Studi Linguistici e Culturali Comparati, Università Ca' Foscari, Venezia, 18.-20. nov. 2015).

DanNet. See wordnet.dk, and andreord.dk (Anders Johannsen)

DDO. Hjorth & Kristensen (2003-2005), Det Danske Sprog- og Litteraturselskab, Copenhagen. Online 2015: ordnet.dk/ddo

Nimb, Sanni & Bolette S. Pedersen (forthcoming): Fra begrebsordbog til sprogteknologisk ressource: verber, semantiske roller og rammer – et pilotstudie. In: *Konferencerapport, Nordiske Studier i Leksikografi 13,* University of Copenhagen.

Nimb, Sanni, Henrik Lorentzen, Liisa Theilgaard, Thomas Troelsgård (2014): *Den Danske*

Begrebsordbog, Det Danske Sprog- og Litteraturselskab, Copenhagen.

Nimb, Sanni, Bolette S. Pedersen, Anna Braasch, Nicolai H. Sørensen and Thomas Troelsgård (2013): Enriching a wordnet from a thesaurus. In *Workshop Proceedings on Lexical Semantic Resources for NLP from the 19th Nordic Conference on Computational Linguistics (NODALIDA)*.Linköping Electronic Conference Proceedings; Volume 85 (ISSN 1650-3740). http://www.ep.liu.se/ecp/088/005/ecp1388005.pdf

Nimb, Sanni & Bolette S. Pedersen (2012): Towards a richer wordnet representation of properties – exploiting semantic and thematic information from thesauri. In: *LREC 2012 Proceedings*. Istanbul, Turkey. http://researchr.org/publication/lrec-2012

Pedersen, Bolette S., Sanni Nimb, Jørg Asmussen, Nicolai Sørensen, Lars Trap-Jensen, Henrik Lorentzen (2009): DanNet – the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary. In: *Language Resources and Evaluation, Computational Linguistics Series*. 10.1007/s10579-009-9092-1. 31 s.http://link.springer.com/article/10.1007%2Fs10579-009-9092-1

Mapping Prehistory

Archaeology and Comparative Linguistics in one interactive map

Mikkel Nørtoft
University of Copenhagen
mattbeets@gmail.com

Two worlds apart

Linguistics and archaeology have a long history of problematic intercommunications. Their different disciplines have often come to conflicting conclusions on the origins of prehistoric peoples, languages, and whether or not migration was a valid explanation for the later historical societies and cultures. Often when attempts were made from one discipline to connect its results with the other, there would be heavy debate and each side would claim that their methods had been misunderstood by the other.

In the past few years an increasing number of scholars on each side have put more effort into understanding and working directly with the other side. Still with much difficulty, but some success has been made[1].

However, the two disciplines are still methodologically difficult to combine, and just as many

1E.g. from the archaeological point of view with David Anthony, James P. Mallory, Marija Gimbutas, Kristian Kristiansen, Rune Iversen, etc. The linguistic side might have been a bit more reserved, but, at least in Copenhagen, scholars like Guus Kroonen, Birgit Anette Olsen, and Thomas Olander are writing articles and establishing research projects with archaeologists.

linguists often have a hard time keeping track of the vast array of archaeological cultures and their cultural traits, many archaeologists have problems keeping score of the many linguistic sound laws and loan words.

Two worlds in one

Therefore, I am working on an interactive map in QGIS to help primarily linguists, but also archaeologists cross-reference their ideas with the other discipline. E.g. the linguist, trying to find a time and place for proposed loan word exchange between two languages, can look in detail at this map and see when and where this exchange is most likely to have taken place. This, in turn, could help to classify the interrelationships of the Indo-European daughter languages, and, e.g. when and how Pre-Proto-Germanic dialects came to Scandinavia.

The map features a time bar that can go from 9999 BC to 9999 AD. Here, detailed and critically revised typology and chronology in archaeology plays an important part for the linguist to see in detail how, where, and when a loan word scenario or a language spread might have taken place.
More worlds to come

The future for such an interactive prehistoric map has great prospects. We could map the known historical words and their change through time, in itself a huge undertaking. The map could thus be used as a kind of geographical-etymological dictionary. We could add phonological sound developments and isoglosses and their corresponding language areas through time. We could map historical events, innovations through time, changing royal and their territories, or military battles through history. We could show prehistoric personal stories illuminated by strontium isotope research. As a webpage or application, the map could be a place for the people in general to learn about the past in an easy and exciting way, and someday we might even incorporate little educational computer games on the map, virtual walks in ancient cities, perhaps with virtual reality compatibility like the Oculus Rift.

First things first

In my presentation, and in the paper, I will focus on the map as a hub for studying archaeological cultures drawn as polygons. I will include a smaller case study using the more detailed point data of archaeological finds in the Late Neolithic period of Denmark. I will go into more detail about the specific features and practical problems of making such a map, and some possible solutions.

I will show the beginnings of a new layer with directions of influence and genomically and archaeologically indicated migrations adding to the understanding of the map’s data.

It is not without problems trying to map prehistory in chronological and geographical detail. When mapping something it becomes very concrete and real, almost like an inarguable statement of reality. Inaccurate data of one archaeological culture can appear directly misinforming when it is put up against more accurate data from another culture. In the talk I will outline the map’s dependancy on strict methodological considerations dealing with these challenges.

I will show how the map can include references to specific relevant articles appearing in a pop-up at the click of a polygon or a point, or links to non-academic articles like Wikipedia and encyclopedias.

The map would thus have the potential to function as a research and reference tool for a wide array

of disciplines and at the same time be used for educational purposes.

Kan digitale utgaver være vitenskapelige utgaver?

Sitering og versjoner

Espen S. Ore
Universitetet I Oslo
e.s.ore@iln.uio.no

1. Problemet

Digitale utgaver kan stå som en løsning når det gjelder å få ut utgaver som vil bli for dyre å publisere på papir. Den siste store trykte konkordansen utgitt i Norge, var Ibsenkonkordansen som kom ut i 1993[1]. Og kanskje den store Ibsenutgaven, Henrik Ibsens skrifter blir den siste store komplette papirutgaven i Norge. Men de digitale utgavene byr ikke bare på nye muligheter, de byr også på nye problemer. For at en vitenskapelig utgave skal ha noen verdi i forskning, må man kunne sitere den og referere til steder i teksten som blir sitert. Da må det være mulig for andre forskere å gå til utgaven og finne den samme teksten. Dette er normalt ikke noe problem når det gjelder papirutgaver, men digitale nettutgaver er ikke så konstante. Dette skyldes flere forhold: det er lett å rette/endre en digital utgave - og for nettutgaver som ikke er bundet til et fysisk medium, er endringer usynlige hvis de ikke blir markert spesielt. Dette problemet la undertegnete frem på JADH2015-konferansen i Kyoto i september 2015 ("Digital Editions and Citing Them"), og denne presentasjonen tar problemet videre og diskuterer muligheter som kom opp under diskusjonen i Kyoto.

2. Noen utgaver

Det har i løpet av de siste 15 årene kommet en god del større digitale utgaver i Norden. Den første, Bergen Electronic Edition of Wittgenstein's Nachlass kom i første omgang ut på CD-ROM og var på denne måten bundet til et fysisk medium, men de senere digitale utgavene har først og fremst vært nettbaserte. For en god del av disse utgavene skjer det vedlikeholdsarbeid og dermed finner vi endringer/nye versjoner av utgavene. Det er varierende hvor langt det er mulig for forskere både å ha oversikt over endringene og å kunne peke til en versjon med håp om at denne versjonen også er tilgjengelig i femtiden.

Ingen av arkivene eller tekstutgavene har løsninger som garanterer at en referanse som for eksempel http://www.ibsen.uio.no/DRVIT_UF%7CUFht.xhtml (peker til De unges Forbund i editert utgave nåværende versjon i Henrik Ibsens Skrifter - HIS) vil peke til samme utgave også i fremtiden. Derimot har flere av dem muligheter til å legge til tilleggsinformasjon om versjon som for noen av utgavene gjør det mulig å finne tilbake til teksten som er referert. For HIS er det i november 2015 versjon 1.1 som er siste versjon. Hvis man ønsker å vise til utgaven fra versjon 1.0, kan man finne den via enkelte omveier som http://www.ibsen01-00.uio.no/DRVIT_UF%7CUFht.xhtml.

Jeg vil i diskusjonen gå gjennom følgende prosjekter som viser noen forskjellige løsninger eller

1 Noreng, Harald, Knut Hofland og Kristin Natvig, *Konkordans over Henrik Ibsens dramaer og dikt,* Universitetsforlaget, Oslo 1993

mangler på løsninger i disse problemene. Jeg har valgt å dele dem inn i tre grupper: digitale arkiver, digitale tekstutgaver og digitale faksimileutgaver:

Digitale arkiver:
MENOTA (norrøne tekster)
eMunch
The Wittgenstein Archives at the University of Bergen

Digitale tekstutgaver:
Ludvig Holbergs Skrifter
Grundtvigs Værker
Søren Kierkegaards Skrifter
Zacharias Topelius Skrifter
Henrik Ibsens Skrifter
Selma Lagerlöf-arkivet
bokselskap.no (NSL)

Digitale faksimileutgaver:
Bokhylla (Nasjonalbiblioteket i Norge)

3. Mulige løsninger?

I eksempelet fra Henrik Ibsens Skrifter over, kunne man se at URIene som brukes kan finnes i en versjonsløs form: http://www.ibsen.uio.no/DRVIT_UF%7CUFht.xhtml og i en form med versjonsinformasjon: http://www.ibsen01-00.uio.no/DRVIT_UF%7CUFht.xhtml. Når det gjelder filstier og lignende i URIer, vil det være desginmessige krav som at man ønsker enklest mulig URI, ikke vil belemre brukeren med informasjon som kan være unødvendig - de fleste som går til Ibsensidene vil bare lese "teksten" uten å tenke på grunnleggende spørsmål om hva som måtte være "teksten". Det er derfor lite aktuelt å beholde kompliserte filstier til forskjellige versjoner og regne med at dette vil kunne ha varighet når for eksempel utgavene flyttes til nye systemer. I stedet må versjonsidentifikasjon kobles til det enkelte elektroniske dokument. Dette er heller ikke alltid trivielt, digitale dokumenter kan være dynamiske, generert i øyeblikket fra en kombinasjon av databaser og programmer, men i utgangspunktet vil jeg holde fast ved denne tanken.

Varig identifikasjon av digitale dokumenter kan gjøres gjennom forskjellige tjenester. Ingen av dem er fullstendig idéelle, men de kan gi et utgangspunkt for videre diskusjon. Uansett er behovet at digitale utgaver bør behandles som arkiver og lagres ved institusjoner som har et evighetsperspektiv som nasjonalbiblioteker eller lignende. Alle versjoner som har vært publisert må bevares og være tilgjengelige. Når en versjon blir publisert, må det være tilgjengelig tilstrekkelig med metadata til at en bruker kan lage varige referanser til denne utgaven.

Her er noen eksempler på mulige verktøy tilgjengelige i Norge i dag:

DOI (Digital Object Identificator) baserer seg på betalte tjenester og registre og brukes først og fremst av kommersielle utgivere, men også av andre.
I Norge har Nasjonalbiblioteket (NB) i sin tid bygget opp en tjeneste for utdeling av URN, men det er usikkert hvor lenge dette vil være tilgjengelig. Imidlertid synes det som om NB bruker URNer internt til varig identifikasjon av materiale de publiserer på nettet, så en mulig hybridløsning ville

være om utgaver kunne utgis/deponeres på NB og slik komme inn i varig lagring/fremhenting. På den annen side har ikke NB noen praksis for versjonering av det de utgir digitalt, så dette er ikke uten videre en komplett løsning.

The Copenhagen Associations Project

Jan-Mathieu Carbon, Mario C.D. Paganini, Stella Skaltsa
The Saxo Institute, University of Copenhagen
fpz408@hum.ku.dk, ncf698@hum.ku.dk, stella@hum.ku.dk

This paper presents the work of the Copenhagen Associations Project (CAP), the first systematic investigation of private associations in Classical Greece, the Hellenistic world, and the Roman East (ca. 500 BC – AD 300). Directed by Professor Vincent Gabrielsen at University of Copenhagen, the principal objective of the project is the compilation of a comprehensive inventory of associations from the Greek-speaking world. By association, we mean privately constituted groups, whether of religious devotees or professionals or family members or still others, which are attested by a variety of ancient cultural artifacts, principally inscriptions, papyri and literary sources. In these scientific regards, the project is highly innovative, since it seeks to marry the study of a multiplicity of evidentiary sources towards a unified goal: understanding the development, the distribution and the modes of operation of private groups and collectivities across the ancient Mediterranean.

To fulfil this objective, an online database has been developed, in which all of the material is systematically analysed with regard to the names of the associations, their constitution, membership, property, rules, etc. Another original component of the project's database is that it has involved the collaboration of the more than two dozen international contributors, whose valuable contribution—coordinated and complemented by the work of our team in Copenhagen—has made the realisation of the project possible. The resulting inventory (CAPI), containing approximately 2000 detailed entries, is a fully searchable electronic resource, which is now in the process of being edited and converted into an open access website.

As the website of the inventory (CAPI) will be published in the first half of 2016, it is expected that a demonstration ("Show & Tell") of the online tools will be possible as part of the presentation. Such a report on the Copenhagen Associations Project and the CAPI at the 1st Annual Meeting of the Association for Digital Humanities in the Nordic Countries (DHN) represents a crucial opportunity to introduce colleagues to a multidisciplinary online resource, developed under the auspices of the University of Copenhagen but within an international framework. It is to be hoped that the challenges of the project's conception and development will thus be useful for other incipient Nordic digital humanities projects, and that the CAP team will likewise learn from these other endeavours as it moves towards future challenges: ensuring the durability, continuity and possible expansion of its online resource.

Coping with disaster: the impact of Stórabóla on abandonment, reoccupation and land use in early 18th century Iceland

Gísli Pálsson
Umeå University
gisli.palsson@umu.se

Scale leads ineluctably to generalization: the particular and the individual is effaced in favour of clear patterns. This tension between the general and the particular has implications for approaches at both ends of the scale: studies of individuals can be difficult to contextualize in terms of broader processes, and studies at scale can often feel out of touch with human experience. Bridging the two scales is eminently achievable, but not without the application of digital tools such as SQL and noSQL data structures that allow the general to be formed *on the fly* from particularities.
This paper demonstrates the value of both relational data structures and network theory in the interpretation of a large demographic collapse in early 18th century Iceland. Between 1707 and 1709 a smallpox epidemic ravaged the country. This outbreak, known as *Stórabóla,* may have contributed to the death of a third of Iceland‘s population. Prior research has demonstrated a range of impacts on Iceland‘s society; the effect on settlement dynamics and land use yet to be fully explored, however. A key omission in earlier work has been the ability to investigate the way broad national social and ownership networks affect the recovery process and land use reorganization on a farm by farm basis. The study is based on a full digitization of *Jarðabók Árna Magnússonar og Páls Vídalín,* a census of the country undertaken both before, during and after the population collapse. By pulling together the entirety of the cens fulfil us, organized through a relational PostgreSQL datastructure, the study is able to identify key agents in the human ecodynamic impact of the smallpox epidemic that would not have surfaced if the study were structured on a county-by-county basis.

Travelling TexTs: A HERA financed five countries project from the point of view of its Finnish team

Viola Parente-Čapková
University of Turku
viocap@utu.fi

In my paper, I want to present the HERA-funded project HERA Project *Travelling TexTs 1790–1914. Transnational Reception of Women’s Writing at the Fringes of Europe* (http://travellingtexts.huygens.knaw.nl/) with its background in the COST Action IS 0901 WWIH (Women Writers in History: Toward a New Understanding of the European Literary Culture; http://www.cost.eu/COST_Actions/isch/IS0901) and in the Women Writers’ Networks (http://www.womenwriters.nl/index.php/Women_writers'_networks). More specifically, I would like to talk about some preliminary results of the Finnish team. The project combines the fields of literary history/reception studies and digital humanities.

The *Travelling TexTs* project traces and compares networks created through women’s writing and its

reception from the perspective of five countries at the fringes of Europe (Norway, Finland, Slovenia, Spain, the Netherlands). It studies the role of women's writing in the transnational literary field during the long 19th century. The research is based on a systematic scrutiny of reception data from large-scale sources (library and booksellers' catalogues, the periodical press), mapping women's participation in this process. In this way, the relation between centre and periphery is questioned from a gendered point of view. Like the earlier activities (the Women Writers' Networks and the COST Action), the *Travelling TexTs* project orbits around a shared digital tool, the database *WomenWriters*. The database is being currently developed into a Virtual Research Environment (VRE), providing not only advanced technology for statistical analysis, charting and visualisation, but also the possibility to work together in the virtual space. The project aims at contributing to the development of new, transnational models of writing the history of European literary culture.
In my contribution, after presenting the project as a whole including its digital tool, I would like to focus on the work done by the Finnish team, which has been working with various sources illuminating women writers' reception in the 19th century Finland. Particularly, I want to talk about the research carried out within the Finnish Historical Newspaper Library (http://digi.kansalliskirjasto.fi/sanomalehti/search) combined with other sources as the *Fennica* , the National Bibliography of Finland (https://fennica.linneanet.fi/vwebv/searchBasic?sk=en_FI), and the Old Collection of the Turku City Library, which is being currently catalogued (http://digi.kirjastot.fi/collections/show/1), and which is rich in women's writing (Turku City Library being one of the Associated Partners of the *Travelling TexTs* project). My particular interest is in the way European peripheries were linked by women's writing in the long 19th century and what kind of questions and issues arises from a comparative analysis of the European fringes, enabled by our digital tool (e.g. comparing the Gustav Cygnaeus Collection, part of the Old Collection of the Turku City Library, to the Hedwig von Radics-Kaltenbrunner library collection in Ljubljana, Slovenia). Working in the VRE and visualising the gendered transnational production-circulation-reception process should enable a kind of a transnational look at the 19th century literary culture that has not been possible so far. The research carried out in the *Travelling TexTs* Project should also lead to new approaches in literary history.

The ELMCIP Project, Electronic Literature Knowledge Base, and Data Visualization Research Outcomes

Scott Rettberg
University of Bergen
scott.rettberg@uib.no

The HERA-funded collaborative research project Electronic Literature as a Model of Creativity and Innovation in Practice (ELMCIP) that took place from 2010-2013 was one of the largest European-funded digital humanities research projects led by a Nordic partner in recent years. Led by PL Professor Scott Rettberg and the Electronic Literature Research Group at the University of Bergen, the project included seven partners from six countries, and resulted in outcomes including books, articles, an online anthology of European digital literature, several films, and an open-access research database. This presentation will both provide an overview of this project and focus specifically on one of its ongoing products, the ELMCIP Electronic Literature Knowledge Base (http://elmcip.net/knowledgebase), and on new types of research outcomes including "distance reading" data-mining and visualization projects that can be derived from it.

The ELMCIP Electronic Literature Knowledge Base is an open-access, online database tracing activity in and around the field of electronic literature and the digital literary arts. Inspired by Ted Nelson's vision of literature broadly understood as "an ongoing system of interconnecting documents" (Nelson 1981: 2/9) the Knowledge Base is collecting and connecting bibliographic information and archival materials about the literary production in this field. As this information is linked and cross-referenced in various fields in the Knowledge Base, the relations between objects and actors in the field of electronic literature become explicit, perceptible, recognizable, and communicable. Together these relations comprise the field. In the Knowledge Base, they are defined through content types that include *authors*, *creative works, critical writing*, *events, organizations*, *publishers*, *teaching resources, databases and archives,* and *platform/software*. The Knowledge Base now includes more than 10,000 cross-referenced records in these primary content types. The project is further part of the Consortium on Electronic Literature (http://eliterature.org/cell/) that is linking our database with other international projects in the field in Europe, the United States, Canada, and Australia, providing mechanisms for shared search and developing documentation standards for electronic literature.

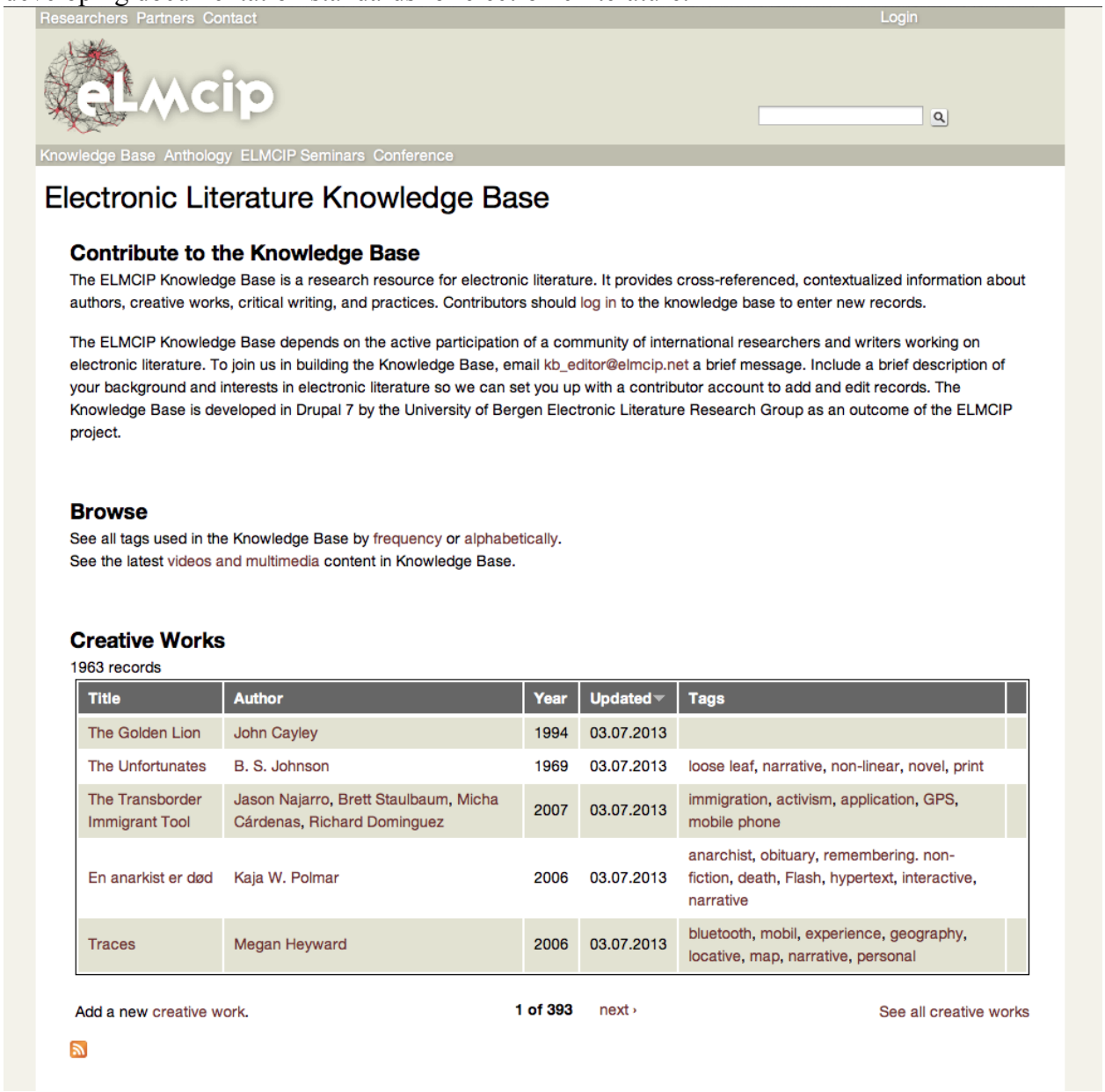


Fig. 1 Front page of the ELMCIP Electronic Literature Knowledge Base

The Knowledge Base is intended to document the electronic literature as a dynamic field of practice, one whose cultural import becomes more comprehensible when the activities of authors, scholars, publications, performances, and exhibitions can be related to each other, in multiple configurations. The Knowledge Base is a platform in which this complex web of relationships can be made visible and available for analysis. Researchers can begin to trace the activities generated or enhanced by a work as it circulates among different reading communities. When a record of a critical article is documented in the Knowledge Base, all the creative works it references are noted, and cross-references then automatically appear on the record for the work itself. Similarly, cross-references are made to every other type of record it touches—when a work by a particular author is entered, a reference automatically appears on that author's page, likewise for works published by a publisher and so forth. The Knowledge Base makes perceptible interactions between human and nonhuman actors, and it documents the diverse range of artistic, scholarly, and pedagogical practices in the field of electronic literature.

An additional application of the Knowledge Base is that as it has been used to aggregate a great deal of information about various objects and actors in the field of electronic literature, it becomes increasingly useful as a resource for doing other types of research based on digital methods. In the University of Bergen Electronic Literature Group, we have begun to do *big data*-style research based on comprehensive reading of changes in trends, themes, technologies and platforms, genres, etc. over time. We have exported specified sets of information from the Knowledge Base and then pulling them into visualization tool-sets, particularly the Gephi network visualization platform, in order to spot and visualize trends and patterns in the field in order to address specific research questions (for example: What is the relationship between platform and genre in digital literature? What works can we consider to be in a “canon” of electronic literature based on citations? What types of theories and themes have been addressed by dissertations in the field?). While this type of research, which Stanford-based scholar Franco Moretti describes as “distant reading,” will never replace close humanistic analysis of literary works it is very useful to develop new perspectives on and contexts for a field which has developed and evolved as quickly as electronic literature in the past several decades. This paper will detail some of the results from this ongoing research project and show how the database serves both an active reference resource and research environment for macroanalysis of the field.

De svenska medeltidsbreven som digital resurs

Sara Risberg
Riksarkivet
sara.risberg@riksarkivet.se

Svenskt Diplomatarium är både en enhet inom Svenska Riksarkivet och ett utgivningsprojekt med lång historia. Grunden för och syftet med vårt arbete är att tillgängliggöra de svenska medeltidsbreven, diplomen, för forskare och allmänhet. Arbetet med textutgåvor av brevmaterialet inleddes redan på 1820-talet och pågår än idag.

På 1990-talet påbörjades ett första digitaliseringsprojekt inom redaktionen för Svenskt Diplomatarium. I detta skede var det lapp- eller kortkatalogen med uppgifter om alla kända svenska medeltidsbrev som var föremål för en digitalisering. Lappkatalogen, som gick under namnet ”Svenskt Diplomatariums Huvudkartotek över medeltidsbreven”, kom att utgöra grunden för det digitala register som även fortsättningsvis kom att benämnas så.

I och med publiceringen av CD-skivor innehållande detta digitala register kunde fler ta del av informationen och forskare lättare på egen hand söka upp enskilda brev av intresse för just deras forskningsändamål. Tack vare ett stort engagemang, såväl från dessa forskare och intresserad allmänhet som Diplomatarieredaktionen och Riksarkivet, kunde det digitala registret publiceras online i början av 2000-talet. I samband med detta kom förkortningen SDHK för Svenskt Diplomatariums Huvudkartotek att etableras som namn.

Nästa steg blev att inkorporera de utgivna texterna i registret. De gamla volymerna skannades, OCR-tolkades och korrekturlästes innan de fördes in och därmed blev sökbara. Vartefter nya häften med brevutgåvor har publicerats, har även dessa texter införts i SDHK. Utgivningstakten idag är att vi vartannat år publicerar ett häfte innehållande ett medeltida års brev. SDHK utökas samtidigt med dessa sökbara brevtexter. Tillgängliggörandet av texter genom kritiska utgåvor är ett tidskrävande filologiskt arbete. För att öka tillgängligheten till diplomen publiceras därför också bilder fortlöpande, så att forskare själva ska kunna läsa och ta del av brevtexterna.

Sedan den första online-publiceringen har SDHK förstås genomgått ett flertal förändringar och utvecklats på många områden. Det senaste året har SDHK förvandlats på ett mer genomgripande sätt. Medan användaren märker förändringen framför allt i postvisningen, som nu har ett helt nytt utseende, har omvandlingen inneburit en stor förändring bakom kulisserna. Idag finns alla data i enskilda xml/tei-filer som indexeras av Riksarkivets söktjänst för att bli sökbara och tillgängliga.

Syftet med detta paper är att presentera SDHK och hur arbetet inom Svenskt Diplomatarium har förändrats under tid. Såväl förr som nu står redaktionen med ena foten i historien och den filologiska grundtanke som legat fast sedan 1800-talet, och den andra foten i dagens och framtidens möjligheter när det gäller att tillgängliggöra vårt kulturarv.

Data-driven text analysis for digital humanities: some thoughts on how and why

Andrew Salway
Language and Language Technology Group, Uni Research, Bergen, Norway
andrew.salway@uni.no

Over the last two decades significant advances in computational power and concomitant techniques for data analysis and visualization have led to a shift in research practice across the natural sciences, with the establishment of data-driven research paradigms. Now the humanities are witnessing a nascent data-driven turn of their own, stimulated by rapidly growing digital archives of cultural resources which include large volumes of text. This paper will reflect on how and why data-driven text analysis methods could and should be used as part of digital humanities research methods, by making particular reference to literary studies and by considering the relevant state-of-the-art technology for natural language processing – including our own work on new technologies.

With the availability of text corpora comprising many millions or billions of words, and the computational power to run inductive algorithms over them, data-driven methods become a feasible and necessary companion to hypothesis-driven approaches. In broad terms, inductive text analysis techniques highlight interesting linguistic patterning in a corpus – for a researcher to then interpret – thus alleviating the problem of "seek and ye shall find" by minimising prior assumptions about the content of the material. Inductive techniques are suited to the exploratory phases of research during

which they can provide researchers with a manageable overview of the content of a corpus in order to stimulate new hypotheses. They may also guide the development of coding schemes, i.e. systematic mappings between countable linguistic forms (e.g. words, phrases, grammatical structures) and concepts related to the phenomena of interest to researchers (e.g. topic, style, framing) which are essential if further automated techniques are to be used to test hypotheses on a large scale.

Approaches to literary and narrative inquiry have traditionally been grounded in the phenomenology of reading (Herman 2005: 125-6) whereby analysts consult their readerly intuitions about a text. In turn, these intuitions find expression in terms and categories deriving from a history of similar engagements by a larger community of expert readers (cf. Toolan 2001: 22). Drawing on these resources, scholars classify literary devices, analyse how they are used in different texts and theorise about how the devices enable readers to understand and engage with narratives. By contrast, the notion of 'distant reading' suggests that keeping a distance from the intricacies of individual works provides new insights by allowing interconnections to become apparent through the analysis of larger collections of works (Moretti 2013).

An interesting approach that blended distant and close reading was taken by Mahlberg in her corpus stylistics work on a Dickens corpus of 4.5 million words (Mahlberg 2007). A computer system identified unusually frequent word clusters compared with a reference corpus of other 19th-century novels. The clusters were assumed to reflect local textual functions and they were manually filtered, organised and interpreted as such, and then used as the basis of stylistic analysis; a similar approach was presented by Herman and Salway (2007), and Salway and Herman (2011). A key point about data-driven approaches is that once an interesting phenomenon has been identified (e.g. a local textual function), its linguistic realisation is already known (e.g. a word cluster that was induced from the text). This means that further data about the distribution of the phenomenon in the corpus, i.e. for testing hypotheses, can be generated automatically without the need for prohibitively expensive manual coding.

The potential utility of large text corpora depends upon the extent to which interesting phenomena manifest in linguistic patterning that is sufficiently regular and frequent that it can be automatically induced. In part there is a fundamental limit since context and sometimes co-text are not accessible to inductive text analysis techniques. There is also a limit due to the state-of-the-art in natural language processing. Whilst insights can be gained using relatively simple techniques (e.g. frequency and keyword lists, n-grams, concordances and collocation data), there seems to be a need for techniques that can induce richer linguistic patterning from corpora. To that end we have adapted a grammar induction algorithm (Solan et al. 2005) and demonstrated how it can be used to induce salient information structures from a variety of corpora (Salway and Touileb 2014; Salway, Touileb and Tvinnereim 2014; Touileb and Salway 2014). The induced structures extend and enrich the overview of corpus content provided by extant techniques. In the current paper we will show and discuss preliminary results from applying our method to a literary corpus, and compare results with Mahlberg (2007).

References

Herman, D. (2005). "Quantitative Methods in Narratology: A Corpus-based Study of Motion Events in Stories". In: J. C. Meister (ed.), *Narratology Beyond Literary Criticism*. Berlin: de Gruyter, pp. 125-49.

Herman, D and Salway, A. (2007). "New Foundations for Narrative Theory: a corpus-based approach." Paper presented at *SSNL International Conference on Narrative*, Washington DC, March 2007.

Mahlberg, M. (2007). "Corpus Stylistics: Bridging the Gap between Linguistic and Literary Studies". In M. Hoey, M. Mahlberg, M. Stubbs, and W. Teubert, *Texts, Discourse and Corpora: Theory and Analysis*. Continuum, pp. 219-46.

Moretti, F. (2013). *Distant Reading*. Verso Books.
Salway, A., and D. Herman (2011). "Digitized Corpora as Theory-Building Resource: New Methods for Narrative Inquiry". In: R. Page and B. Thomas (eds.) *New Narratives: Theory and Practice*. University of Nebraska Press.
Salway, A. and Touileb, S. (2014). "Applying grammar induction to text mining". *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp.712-717.
Salway, A., Touileb, S. and Tvinnereim, E. (2014). "Inducing Information Structures for Data-driven Text Analysis". *Proceedings ACL workshop on Language Technologies and Computational Social Science*.
Solan, Z., D. Horn, E. Ruppin, and S. Edelman (2005). "Unsupervised learning of natural languages". *PNAS* 102 (33), 11629-11634.
Toolan, M. (2001). *Narrative: A Critical Linguistic Introduction*. London: Routledge.
Touileb, S. and Salway, A. (2014). "Constructions: a new unit of analysis for corpus-based discourse analysis". *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation* (PACLIC 28) eds. W. Aroonmanakun, P. Boonkwan, and T. Supnithi.

Swedish Women Online (SWO) from the Middle Ages to the Present

A Biographical Dictionary of Swedish Women, in Swedish and English

"The evidence for the inferior intellectual capacity of women is so obvious and overwhelming that anyone who can lightly set it aside must be defective in their *attitude* to evidence […]." (David Stove 1990)

Lisbeth Larsson, Maria Sjöberg, Lars Borin
University of Gotheburg
lisbeth.larsson@lir.gu.se, maria.sjoberg@history.gu.se, lars.borin@gu.se

Why are there no female artists, philosophers, scientists, politicians, or authors? This question surfaces every time the history of art, literature, medicine, and philosophy are to be written. One important reason is that writing history is based on the existing biographical dictionaries where the principle of selection is gender biased. Only five to ten percent of the entries in the large Swedish biographical dictionaries are women. Therefore, if writing history is to become less misleading, then the very point of departure – the source material – has to be revised.

As Jenny Teichman says in her response to David Stove (above): "Stove treats the supposed connection between lack of achievement and lack of capacity as on par with the connection between achievement and capacity. But this is a mistake." (Teichman 2001: 150)

The paper presents a project with aim to create an online Biographical Dictionary of Swedish Women – hereafter abbreviated as SWO – which would be linked to other online digital resources and works of reference dealing with Swedish women; this on-line dictionary will itself instigate and contribute to further research and scholarly work.

Theoretical reflections

The women who are represented in the existing biographical dictionaries tend to have been married

to prominent men or to have been active in various cultural professions, for example as actors, singers, or writers. This kind of representation mirrors the male hegemonic view on what constitutes an important woman rather than being a real assessment of the impact of women on the development of society. In order to obtain a realistic picture of women's role in this development, prominent women in the areas of politics, science, popular movements, etc., must also be represented and given a biography even though they may not, for various reasons, 'have made' as big 'a name for themselves' as their male colleagues (cf. Steedman 2001). Women's unpaid work, for example, was of great importance to the stability of society for a long time. Women of various professions executed their work in the shadow of a man and/or together with a man with the result that achievements that might have been considered as having being accomplished by women were not acknowledged as such. Women's entrance onto the public scene also needs to be rendered visible, even though it took place at a later point in time than that of men.

Thus, in order for women to be included in a national biography, the SBL's criterion – that those with biographies must have "made a name for themselves within all areas of society" – needs to be interpreted differently than the established and male based way. The historical context must be taken into consideration (Derrida 1996, Steedman 2001, Ricœur 2005).

Method

In accordance with the results of a pilot project, led by Lisbeth Larsson, the SWO will be using the following criteria for selection: (a) women who have connections to Sweden and have contributed to the development of society in general and/or made efforts internationally and/or at a regional level; (b) historical women of significance who have lived in Sweden; (c) women who were pioneers in different areas; (d) women of significance in the struggle for gender equality; (e) women of regional significance; and (f) women with a connection to Sweden who have made contributions of international significance. Just like SBL, the concept of Sweden and Swedish in SWO follows the historical realities; Finland prior to 1809 will therefore be included. The technical solutions will be developed in collaboration with Språkbanken (the Swedish Language Bank <http://spraakbanken.gu.se/eng>), meaning that the work on SWO will be able to both draw upon and inform the digital humanities research infrastructure being developed under the auspices of SWE-CLARIN <https://sweclarin.se/eng/home>, where Språkbanken is the coordinating partner.

Expected results

SWO will offer a great number of different search options, such as dates, professions, activities, organizations, political affiliations, and other criteria. With many and diverse search and visualization options, it will be possible for the user to quickly find single pieces of information as well as to make more systematic searches in order to focus upon, for example, professional groups, political networks, time periods, and geographical areas.

Codicologist vs. Computer: How can Machine Learning Support Research in Manuscript Studies?

Beeke Stegmann, Anders Trærup Johannsen
University of Copenhagen
beeke.stegmann@gmail.com

Manuscripts (handwritten books) in collections worldwide often appear to the observer quite differently today than when they were originally written. Most of them have been rebound, of course, but in many cases more substantive changes have occurred. For instance, Árni Magnússon (1663-1730), a famous collector of Icelandic manuscripts, frequently took parts of different manuscripts that he deemed related and combined them into new volumes. Discovering the original contexts of these rearranged manuscript parts is an important element in understanding the history of a manuscript as well as of the whole collection. Often the formerly-neighboring manuscript parts have been lost or reside in different libraries. For the Icelandic manuscripts from Árni Magnússon's collection, we are fortunate enough to have most of the pieces of the puzzle.

In this paper we investigate how statistical methods can assist codicologists (manuscript scholars) in performing this reconstruction of original manuscripts. We focus on the task of identifying former parts of the same manuscript books based on physical characteristics, such as paper size, style of hand writing, page decorations, etc. We test manual reconstructions of former books against computer-created clusters of manuscript parts with the aim of exploring the possibilities of automating the reconstruction process. This establishes the feasibility of the approach: can the automated analysis, in some cases, rival a more labor-intensive manual process? A related benefit of the automatic analysis is that we can inspect the statistical models and see what factors (e.g. paper size, hand writing, decorations etc.) were most influential in the model's decision to group two manuscript parts together.

Our corpus consists of 118 XML files, containing information about 237 manuscript parts, their characteristics and physical context in different time periods. The data is encoded in TEI P5-conformant XML and was collected by a trained codicologist (the main author of the paper). Four of the 118 files are so-called authority files, encoding meta-information on the former shapes of the manuscripts as well as scribes, owners and other persons that played a role in the production and provenance of the books. All manuscripts that are included in the corpus belong to the Arnamagnæan Manuscript Collection (housed in Copenhagen and Reykjavík). Of those, only paper manuscripts from the period 1601-1730 are considered that are known to have been physically changed. In total there are 155 manuscript parts which have full information on the context of the period under scrutiny, i.e. when they entered Árni Magnússon's collection.

We propose two types of statistical analyses: supervised manuscript book reconstruction and outlier detection. Both make use of the authority files, in addition to information about the physical characteristics of the manuscript parts. In the manuscript-book-reconstruction task, we train machine learning models to predict manuscript book membership. The setup is as follows. Given a set of already established manuscript books (established by the main author), along with feature representations of their constituent manuscript parts, we wish to predict to which manuscript book a new and unseen manuscript part belongs, if any. The performance of the predictive classifier is tested in a leave-one-out cross-validation setup, in which *n-1* labeled instances train the model and *1* instance is used for test. This process is repeated *n* times until all instances have been used for

testing. For outlier detection, we analyze the manual grouping of manuscripts, flagging any manuscript parts or proposed books that markedly deviate from the norm. This process uncovered, among others, the fact that more manuscripts than previously thought were already altered by the time they came into the possession of Árni Magnússon.

Since digitized descriptions only exist for a smaller subset of the manuscript parts in the collection, a fully unsupervised clustering is currently not possible. However, for the data available, the automated approaches still yield valuable results, especially with regards to outlier detection and troubleshooting. Moreover, being able to inspect the statistical importance of individual features for manuscript reconstruction is helpful for optimizing the traditional research methods. Thus, the statistic analyses performed provide a useful tool *in combination* with manual manuscript reconstruction.

Project presentation: The Jihadi Document Repository

Anne Stenersen, Erik Skare[1]
Norwegian Defence Research Establishment (FFI)/University of Oslo
anne.stenersen@ffi.no, erik.skare@gmail.com

The *Jihadi Document Repository* is a joint project between the University of Oslo and the Norwegian Defence Research Establishment (FFI). The aim is to create a digital collection of Arabic and English primary sources related to the study of jihadism. "Jihadism" in this context refers to the militant Islamist ideology that emanated from the Arab-based foreign fighter movement of the Afghan-Soviet war in 1979-1989.[2]

For many years, researchers at the Norwegian Defence Research Establishment (FFI) have collected jihadi primary sources such as journals, books, leaflets and official communiqués. The collection includes scanned copies of documents dating back to the early 1990s, but a majority of the documents are electronic sources downloaded from the Internet from as early as 2002-2003. Many of these electronic sources are no longer accessible today, making FFI's collection a unique and valuable source for academic research on jihadism.

So far, the FFI's collection of jihadi documents have been scattered on the computers of individual researchers, or in shared folders at the FFI with limited access to outsiders. The aim of establishing the Jihadi Document Repository is three-fold: 1) To gather all of FFI's primary sources in one place; 2) To make the collection searchable; and 3) To share the collection with students and researchers world-wide. By establishing non-commercialism as one of the repository's core tenets, we hope to encourage transnational scholarly cooperation in jihadist research and source gathering.

As with any archive, there are several challenges. A substantial amount of the documents in the repository promote, encourage and, to some extent, facilitate violence. Some of the documents

1Anne Stenersen (M.Phil., PhD) is a research fellow at the Norwegian Defence Research Establishment (FFI). Erik Skare (M.Phil.) is a research assistant at the Norwegian Defence Research Establishment and at the Department of Culture Studies and Oriental Languages (IKOS) at the University of Oslo.

2For a well-informed account of the rise of the Arab-Afghan movement and al-Qaida, consult Camille Tawil, *Brothers In Arms: The Story of Al-Qa'ida and the Arab Jihadists*, trans. Robin Bray (London: Saqi Books, 2011).

contain manuals and recipes on how to conduct military operations, how to use light and heavy weaponry and how to make bombs. In order to prevent abuse of the repository's content, the University of Oslo and the FFI are currently discussing whether to restrict access to pre-approved users (students and scholars) who have to submit an application stating their full name, institutional affiliation and research interests in order to gain access to the repository. In addition, we have chosen to watermark all the documents in the repository in order to prevent commercial actors from selling the sources to a third-party for profit.

Another challenge is to make the archive an attractive and dynamic site that will be used by students and researchers alike. So far we have taken certain measures such as 'metadatafying' all the material in order to lay the groundwork for an optimal search function. Furthermore, we will provide short introductions to each category of primary sources, and create hyperlinks throughout the archive in order to spur the will to explore the material in the repository. In addition we will provide a "latest update" bar on the front page, giving visitors an immediate overview of new content uploaded to the repository.

The official web page of the Jihadi Document Repository will be launched in the spring of 2016, and is FFI's first Digital Humanities project.

Locus Classicus

Tue Emil Öhler Søvsø, Mitchell John Seaton
University of Copenhagen
tues@hum.ku.dk

Locus Classicus (*LC*) is an attempt to bridge the gap between the readers of today and the texts of the past. It focuses on the core activity in Greek and Latin philology - close and concentrated reading - and puts this activity at the centre of an interactive and interdisciplinary portal offering open access to Greek and Latin texts from all periods of these languages' history.

The project builds on a wide range of the digital resources developed during the past decades of digital philological scholarship, but aims to develop them further by collecting them in a digital environment focused exclusively on the reading and interpretation of texts. In doing so the questions we want to address are how we read and work with texts – as students, as teachers and as researchers – and how we can best organize, use and share the knowledge we accumulate in this process.

The backbone of the portal is a reading application allowing the user to navigate easily between: (i) a reading-mode showing a clean Greek/Latin text (and/or its translation) and (ii) an analysis mode offering both text, translation and commentaries on a given sentence in the text. This allows the reader to navigate between text, translation and commentary without losing sight of the sentence he/she was reading.
By hiding all reading aids until the user actively chooses to have them shown this application offers a reading environment which encourages the kind of close reading so essential to textual scholarship. Once it has been activated, however, the analysis mode offers an ideal platform for introducing the reader to a wide range of digital aids and resources without intruding on the user's concentrated reading. Potentially, *LC* can therefore serve both as a reading tool and as an introduction to modern research and digital methods.
A prototype of this reading environment has recently been produced (see www.locus-classicus.org

and attached video presentation) and over the next few years we will continue developing the application and its text corpus. Parallel to these efforts a series of user interviews and tests will be conducted providing an empirical basis for the selection of further features to be offered by the portal.
At present we intend to include three kinds of information in the analysis mode: (i) traditional reading aids such as dictionaries and commentaries, (ii) user- and auto-generated bibliographies and (iii) links to digital research tools.
LC will mainly rely on the digital Greek and Latin dictionaries already available, in particular the *Alpheios* cursor tool (http://alpheios.net/). Our efforts will instead be focused on the production of commentary resources, especially collaborative commentaries. The aim is to allow users to comment directly on the text and evaluate the comments of others but also to develop a way of including the work done within dedicated collaborative commentary projects such as *Cyrus' Paradise* (http://www.cyropaedia.org/).X
By including a bibliography in the analysis mode we hope to develop a supplement to traditional research bibliographies. *LC* aims to create a bibliography organized around the source texts. It will be generated both through user-suggested references and a collaboration with scholarly publishing houses offering them the opportunity to advertise new publications via a labelling system. The publisher simply adds a list of passages to the metadata of the publication, and the publication is then automatically included in the bibliography generated for those passages at *LC*.
LC will rely heavily on content drawn from other digital resources but will present it in a different form and environment. We will thus collaborate with other major digitization projects such as *The Perseus Project* (http://www.perseus.tufts.edu/hopper/) and *The Open Greek and Latin Project* (http://www.dh.uni-leipzig.de/wo/projects/open-greek-and-latin-project/) as well as sites offering digital critical editions. As part of this collaboration it is essential to offer easy access to the original sources of the content and the functionalities offered by these sites. By supplementing these links with links to further resources such as prosopographical and geographical databases we hope to be able to serve as an entry point into the world of digital philology. As with the bibliography, the guiding principle in offering this information will be the relevance of a given link to the passage currently read by the user.X

Combining all of these elements our ambition is to create a new kind of reading, teaching and research environment. A place where people go to read and reflect on the textual remains of our past.

SWE-CLARIN – the Swedish CLARIN project – aims and activities

Nina Tahmasebi, Lars Borin, Caspar Jordan, Stefan Ekman
University of Gothenburg
nina.tahmasebi@gu.se, lars.borin@gu.se, caspar.jordan@snd.gu.se

CLARIN (Common Language Resources and Technology Infrastructure) is a European Research Infrastructure Consortium (ERIC), an ESFRI (European Strategy Forum on Research Infrastructures) initiative which aims at (a) making extensive language-based materials available as primary research data to the Humanities and Social Sciences (HSS) research communities, (b) offering state-of-the-art language technology (LT) as an e-research tool for this purpose.

SWE-CLARIN as the Swedish CLARIN node was established in 2015 with funding from the Swedish Research Council by a consortium consisting of 9 members, representing the Swedish academic community as well as public cultural heritage institutions. The academic members are well balanced over the LT field, covering existing and possible research areas and user groups, and the cultural heritage institutions provide access to many of the possible language-based materials of interest to the users. SWE-CLARIN is coordinated by Språkbanken at the University of Gothenburg.

From the very beginning, the vision in SWE-CLARIN has been to establish good relations to the HSS fields and open the door for all researchers who wish to work with digital humanities research using text and speech as primary research data. To avoid being a project by language technologists for linguists, our aim is to include the intended HSS users in the process as early as possible. Our preferred way of doing this has been for our partners to establish small pilot projects with at least one member from the HSS field and at least one consortium member, together formulating a research question the addressing of which requires working with large language-based materials. Ideally, the collaboration should additionally always include a data owner, a person or persons representing the institution where the text or speech data is kept – typically a cultural-heritage institution.

The pilot projects aim to spread the word of SWE-CLARIN, show the potential of using language technology in digital humanities research, create a user base for the tools and resources developed within the project, and last but not least, having this development being informed by input from users in the earliest possible stages of the project. Some pilot projects are already underway and deal with the history of public discourse on rhetoric; history of gender and work; and using Wikipedia's revision history to investigate language change and sociolinguistic networks.

In addition to the pilot projects, we have arranged workshops and user days and published newsletters. The workshops held so far have been on topics such as: general introduction to SWE-CLARIN, our tools and resources; historical resources and tools; making cultural heritage text data available for research; and HSS research on digitized speech data, such as those of the Swedish Media Archive. For the coming year, we plan a series of workshops called "Swe-Clarin on tour" where the Korp corpus infrastructure will be used to explore previously unexplored materials in a hands-on manner, giving researchers of LT and HSS the opportunity to meet and discuss research questions and the potentials of using LT for digital humanities. The experience from working with researchers from the HSS field will help reveal the limitations of existing tools and hopefully also engender general methodological discussion, thus setting the stage for future development of tools more appropriate for DH research.

To disseminate our results and ongoing research including the pilot projects we publish a monthly newsletter online at <https://sweclarin.se/eng/home> that during this first year has allowed our partners to introduce themselves. During the second year, the focus will be on the pilot projects. Each newsletter also has a calendar where upcoming events are listed.

We encourage contact if you are interested in any of our resources, conducting a pilot study with us or if you have any ideas or questions regarding digital humanities research with respect to language technology and resources.

MANART – en databas som gör konsthistoriens dynamik *manifest*

Mette Tjell
University of Gotheburg
mette.tjell@sprak.gu.se

Kan digitalisering av arkivmaterial påverka vår perception av konsthistorien? Vad händer i övergången från objektens materialitet till virtuella faksimile, placerade sida vid sida i en databas? Vilka forskningsmöjligheter öppnas när hierarkiska relationer på detta sätt plattas ut? Viljan att svara på dessa frågor ligger till grund för projektet MANART, en databas över manifest producerade inom konst och litteratur från slutet av 1800-talet och fram till idag. Projektet, som initierades i 2012 av Viviana Birolli (Paris 1, EHESS), Camille Bloomfield (Paris III, Paris 13), Audrey Ziane (Aix-Marseille Université, CNRS) och undertecknad[1], upptogs i 2014 i den franska, statliga forskningsstrukturen Laboratoire d'Excellence Création Arts Patrimoine (Labex CAP), med vars finansiella stöd projektet vidareutvecklas[2].

Avantgardestudier har sedan 80-talet växt till ett betyande och internationellt forskningsfält (nämnas kan bl a. Bürger, 1984; L. Somigli, 2003; M. Puchner, 2006), men hittills har de verktyg som finns att tillgå inom *digital humanities* bara tillämpats i begränsad omfattning (se ex. Pictorana.be). Möjligen är förklaringen att statistisk forskning kan verka oförenligt med avantgardets fluktuerande diskurser, men då har man inte tagit hänsyn till den stora plasticitet som digitala verktyg erbjuder och som, rätt använda, snarare kan synliggöra deras mångfald. För att belysa detta kommer konferensbidraget att fokusera på databasens utformning och potential att kasta nytt ljus på vedertagna idéer om avant-gardediskursernas mekanismer, något som även kommer att illustreras genom flera exempel.

MANART innehåller i nuläget ca 750 manifest, främst från Europa och USA, men antalet utökas kontinuerligt. Per definition är databasen i ständig expansion, då det inkluderar alla geografiska fält och konstnärliga discipliner (måleri, skulptur, samtida konst, litteratur, dans, teater, film, etc.). Databasens utveckling bygger följaktligen på aktiva och internationella samarbeten mellan forskare inom skilda intresseområden och vars arbete i en gemensam plattform skapar värdefulla synergieffekter.

För att möta komplexiteten i manifestgenren och betydelsen av dess paratextuella aspekter, krävs för varje kartlagt manifest information om dess produktion och reception (publikationsår, land, publikationsform, bibliografiska och kritiska referenser, etc.). På detta vis öppnar databasen för möjligheten att ge ett statistiskt underlag i en rad frågeställningar av både kvantitativ och kvalitativ karaktär. Forskning som hittills bedrivits utifrån den aktuella versionen av databasen, och vars resultat presenteras i konferensbidraget, är bland annat huruvida manifestgenren haft en storhetstid och hur markerad den i så fall har varit. Är manifesten alltid kollektiva? Om inte, hur uttrycks spänningen mellan det individuella och det kollektiva i manifestets form och diskurs? Andra

1Startpunkten för projektet var ett symposium vid Centre de recherches sur les arts et le langage (CRAL), EHESS, organiserat av V. Birolli och M. Tjell. Titel på symposiet : « Le manifeste artistique: un genre collectif à l'ère de la singularité » : http://www.fabula.org/actualites/le-manifeste-artistique-un-genre-collectif-a-l-ere-de-la-singularite_47914.php
Symposieakterna publicerades i tidskriftsnumret « MANIFESTE/S », *Études littéraires*, vol. 44.3, Université de Laval, Québec, 2013 : http://www.etudes-litteraires.ulaval.ca/manifestes/
2Hemsida: http://labexcap.fr/micro-projets/manart-les-manifestes-artistiques-et-litteraires-au-xxe-siecle/

möjliga frågor är naturligtvis hur den manifestära traditionen ser ut från ett land till ett annat – och från ett konstnärligt område till ett annat – liksom hur en viss strömning har förändrats över dessa gränser. Databasen lämpar sig dessutom väl till studiet av genrens diakroniska konstruktion, i skärningspunkten mellan konstnärernas textproduktion och dess reception av konsthistoriker.

Ambitionen med MANART är vidare att, i samarbete med relevanta konstinstitutioner, bygga upp ett digitalt arkiv av konsthistoriens diskursiva samlingar och därmed tillgängligöra texter som annars är svåråtkomliga. Att kunna återge manifestens visuella aspekt är just en av de stora vinsterna med digitala verktyg, som på så vis kan ge en mer komplex bild av texternas slagkraft än som är fallet med dagens antologier i bokform där manifesten huvudsakligen reproduceras som text (M. A. Caws, 2000 ; C. Harrison & P. Wood 1993 ; K. Stiles & P. Selz 2012, etc.). Denna visuella dimension gör det bland annat möjligt att lyfta fram manifestets grafiska utveckling över tid, från avant-gardeperioden och fram till idag

Databasen har redan konverterats till XML-format och kommer under våren 2016 att publiceras på en websida där forskare både kan utföra avancerade sökningar och bidra med ny information. Sökresultaten kommer att visualisera reproduktioner av manifesten och möjliggöra både nedladdning och generering av diagram. Därmed tas första steget mot att göra Manart till ett verktyg *open access* för ett internationellt nätverk och bidra till ytterligare dynamik inom ett högst levande forskningsfält.

Källförteckning

Birolli, V. & Tjell, M, ”MANIFESTE/S”, *Études littéraires*, vol. 44.3, 2013.

Bürger, P., *Theory of the Avant-Garde*, trad. de M. Shaw, préface par J. Schulte-Sasse, Mineapolis, University of Minnesota Press, 1984 [1974].

Caws, M. A. (ed.), *Manifesto: a century of isms*, Lincoln, University of Neraska press, 2001.

Danchev, A. (ed.), 110 Artists’ Manifestos. From the Futurists to the Stuckists, Londres, Penguin, 2011.

Harrison, C. & Wood, P. (ed.), *Art in theory, 1900-1990*, Oxford, Blackwell, 1993.

Pictoriana.be, Portail consacré aux écrits d’artistes, Université Libre de Bruxelles.

Puchner, M., *Poetry of the Revolution. Marx, Manifestos and the Avant-Gardes*, Princeton/Oxford, Princeton University Press, 2006.

Somigli, L., *Legitimizing the Artist: Manifesto Writing and European Modernism, 1885-1915*, Toronto, Buffalo, London, University of Toronto Press, 2003.

Stiles K. & Selz P (ed.), *Theories and documents of contemporary art: a sourcebook of artist’s writings*, London, University of California press, 1996.

Teaching Digital Humanities at the Faculty of Arts at the University of Helsinki

Mikko Tolonen, Maija Paavolainen
University of Helsinki
mikko.tolonen@helsinki.fi, maija.paavolainen@helsinki.fi

Eetu Mäkelä
Aalto University
eetu.makela@iki.fi

The idea of this paper is to put forward a case how teaching of digital humanities is envisioned and currently put to practice at the Faculty of Arts at the University of Helsinki. The paper is co-authored by a multidisciplinary team responsible for implementing digital humanities teaching at HY with background in the humanities, computer science and the university library. Setting up DH teaching at HY is an ongoing process.

Our approach to DH can be described as pragmatic, we operate with working definitions without an attempt to problematize particular aspects of them: DH can be many things. This also means that we have designed one common study block for the arts faculty that benefits from – but also respects – the multifaceted nature of the humanities. Focus is on hands-on projects, such as hackathons and other forms of multidisciplinary collaboration. The starting point is to work with relevant humanities research questions, materials and data. The syllabus is also founded on pragmatic basis by building on a list of existing courses taught elsewhere within the faculty and beyond.

Teaching philosophy of the program follows these same principles. Students are introduced with real problems, methods and coding to some degree from the very beginning. Emphasis is put also on questions that are pressing also in research: good example of this is the implementation of best practices of open science and relevance of work ethics and in many cases of the hard question of cleaning and improving the quality of (often digitized) humanities data before reaching conclusions in the form of visualizations.

Students are also familiarizing themselves with possible future research data by mapping the heritage collections of Finnish GLAM organizations and learning to evaluate them from the viewpoints of the quality of documentation, machine readability and reuse. The Open Culture movement principles guide the licensing and copyright considerations concerning heritage material and help students to conceptualize relevant research ethical topics.

Hands-on teaching begins already during the introduction to digital humanities course in a form of a mini research project. Different aspects of group work in the form of tools and documentation are emphasized in all of the teaching. For example, interactive tools documenting discussions in class are used throughout the courses.One important part is also to bring in MA students with humanities and computer science background together in a classroom as often as possible.

In general, it needs to be underlined that collaboration far beyond the link between humanities and computer science is one of the core elements of DH at HY. Different joint efforts, multidisciplinary work and conscious practice of open culture and science will lower the barrier between different research projects on similar data that will enable DH to deliver what it has promised. This is also a good way of mainstreaming research groups in the humanities.

Open access, and open science are a core principle of all of our DH activities, we will explain how this is implemented in the teaching. What is important in this change towards a more open working culture is borrowing of the best practices from other fields of science. Although methods from natural science cannot be implemented as such in the humanities, we still have much to learn from the ideal of open, transparent and reproducible workflows that are these days a mainstream feature in bioinformatics, for example. At the same time, it is equally important to keep the emphasis on what is particular in the humanities and digital humanities, which is the heterogeneous nature of the research questions and humanities
data.

DH teaching ties in with other digital humanities infrastructural processes in Helsinki, which will be also explained during the talk. These processes include: (1) founding a DH centre (HELDIG), (2) formal infrastructural collaboration outside the University (with Aalto University in particular) and (3) integrating DH research within the University, for example, by organizing different conferences and events and running a Digital Humanities Research Seminar for DH doctoral students and other researchers working on DH (link to many of the DH activities at HY: http://blogs.helsinki.fi/mstolone/) It needs to be underlined that digital humanities is an emerging field at the University of Helsinki.

Conjuring up the Artist from the Archives: Ivar Arosenius

Digitization and Coordination of Archives for Enhanced Accessibility and Research

Jonathan K Westin
University of Gotheburg
jonathan.westin@gu.se

How can our understanding of an artist be deepened and developed through digital materials and methods? How can we develop tools not only for exploring the possibilities, but also for a better understanding of previous practices in conjuring up, modifying and curating artists and works of art in museum exhibitions, publications and studies? What ideological and practical considerations and presuppositions have governed the presentations that have formed the artist for the public consideration?

These are questions we have started to investigate through a three-year project about Ivar Arosenius, a Swedish artist and writer. His main body of work was produced during the last few years that led up to his untimely death in early 1909, only 30 years of age and within months of his big breakthrough. During the subsequent years and decades, his substantial production earned him recognition posthumously both nationally and internationally, and today he is viewed as one of the most renowned Swedish artists.

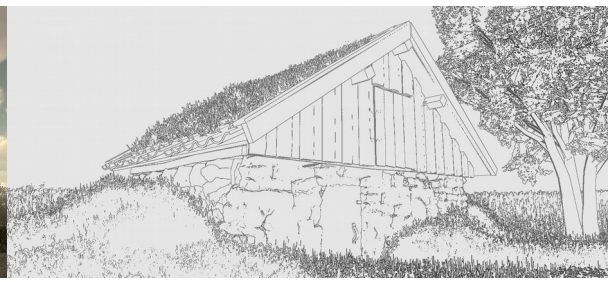
Built on the foundation of a deposition of manuscripts obtained from the Royal Society of Arts and Sciences in Gothenburg in 2001, the Ivar Arosenius Archive at the university library in Gothenburg has over the years swelled both through additional donations and accessions. The archive, more a collection of documents associated with Ivar Arosenius than a personal archive, encompasses both

originals and copies dating from 1896 to 2012. In addition to the many photographs, letters and legal documents pertaining to Ivar and his closest relatives, the archive also includes an accumulation of exhibition catalogues and newspaper clippings.

The Ivar Arosenius project builds upon this archive by developing a platform for collecting the digitized material from several additional archives, both public and private, into a whole, where both well-known works and works that have received less attention, as well as hitherto hidden objects – drafts, letters, note-books, photographs and an array of documents – are made available so as to allow scholars and the public to view, filter, search and combine the entirety in new ways, according to motifs, periods, persons and localities.

The development will entail a number of studies of what knowledge and aspects can be added through different technological developments, as well as what knowledge and aspects are lost or threatened in a digitization process. The presentation will focus on two advanced studies; the processes of translating a physical archive into a digital, and methods through which to give body, context, and affect back to a digitized material. In the former study we follow the material as it travels from the manuscript vault into the digitization studio, mapping all the actants involved in shedding it of its physicality. The latter study explores material pertaining to Arosenius' home, lost to a fire in the early seventies after decades of neglect. Using the archive as a source, a virtual model is constructed in Unity where the connection between artist, art and place is investigated to catch the way Arosenius has translated his surroundings.

With this interactive reconstruction we aim to construct a synthesis of a heterogeneous and sometimes conflicting material that can be used both as an access-point to the life of Ivar Arosenius and his art, and as a repository: built on a source material consisting of archival photos, local stories and historic maps, paintings, 3d-scanned artefacts, sound recordings, and inventories of both the belongings of the artist and his family, and of the vegetation on his lands, the digital construction will be a knowledge-model containing all the material pertaining to this part of the artists' life. As such, of central interest for the study is how to communicate interpretative practices to the user, balancing an incomplete source material with the need to create a space that can inspire affect.



The project involves a number of departments and divisions at the University of Gothenburg as well as the Swedish National Museum in Stockholm, and the Museum of Art in Gothenburg.

One for all - How to transcribe Icelandic manuscripts for scientific research

Ludger Zeevaert
The Árni Magnússon Institute for Icelandic Studies
ludger@hi.is

The Icelandic manuscripts are an indispensable source for the investigation of Icelandic language and literature. However, their scientific exploitation still poses a challenge for the application of methods typically used in digital humanities. Some peculiarities of pre-modern Icelandic writing make a diligent transcription rather time-consuming, and optical character recognition is not yet ready for use with medieval or early modern Icelandic script. Obstacles for digital methods are for example allographic variation (round/long *s*, *r* and *r*-rotunda, *f* and insular *f* etc.) and the, in comparison to continental manuscripts, excessive use of abbreviations.

Different research questions call for different types of text-representation. For historical, literary or folklore studies normalised text editions (or English translations) may be sufficient, but linguistic, philological or paleographic studies are dependent on a more accurat rendering of the manuscript texts (see Zeevaert 2014). The compilation of an automatically lemmatised corpus requires orthographically normalised transcriptions, whereas morphological and phonological variation can only be studied on a more manuscript-near representation of the spelling.

To avoid the multiple manual transcribing of the same manuscript on different levels of accuracy it would be advantageous if different representations of one text needed for different research questions could be produced automatically from one transcription that can be prepared with an affordable amount of time. The optimal approach would go out from a maximally accurate transcription that could then be downgraded step by step. In a project with limited financial and temporal resources, however, this is not feasible due to the amount of work and time needed for such a transcription. In the project "Gullskinna. Postmedieval transmission and reception of a lost medieval parchment-codex" we therefore go out from a semi-diplomatic transcription of different manuscripts of *Njáls saga* which can be used as a starting point for an electronic edition of the text. To tackle linguistic questions morphological and syntactic annotations are added.

Although a more detailed level of transcription is not needed for our research it would be useful for other scholars to have at their disposal a transcription that renders allographic variation and abbreviation signs. Such a type-facsimile transcription is for example favoured by the *Medieval nordic text archive* (www.menota.org). We therefore decided to find out whether our semi-diplomatic transcription could be used to compile a type-facsimile transcription automatically. Our approach is based on the consideration that the medieval and early modern Icelandic writing system is, in difference to the modern Scandinavian practice, not simply an alphabetic script but a combination of a logographic, a syllabic and an alphabetic system (for examples see figures 1-3). This system is not completely random but based on rules that can be shown on manuscripts like *AM 136 fol.* from around 1640, the main manuscript in the *Gullskinna*-project:

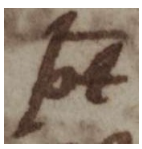


Fig. 1: Example for a logogram from manuscript *AM 136 fol.*: the abbreviation for *því að* 'because'

Fig. 2: Example for the abbreviation of combinations of letters from manuscript *AM 136 fol.*: the abbreviation for *-ur* in the word *suður* (ſud*ur*) 'south' (in the manuscript no distinction is made between <ð> and <d>).

Fig. 3: Example for allographs from manuscript *AM 136 fol.*: different *s*-forms in different positions of the word *síns* (ſynꝫ) 'his/her' (the dashes above the *n* are part of the letter *y*).

Our semi-diplomatic transcription follows the orthography in the manuscript but expands abbreviations and reduces the allographic variation to a certain amount. In a first step words that are represented by logograms in the manuscript have to be replaced on the basis of a list with logograms and their expansions. In a second step letters and combinations of letters rendered by superscript abbreviation signs have to be transferred to their original shape. Expanded abbreviations are marked with a certain tag in the transcription, and even here a list of letter-combinations and their abbreviations has to be used. In a third step certain graphemes at certain positions in the word have to be replaced with the allographs used in the manuscript. This has to be done according to certain rules.

Literature:

AM 136 fol. *(Njáls saga)*. Stofnun Árna Magnússonar í íslenskum fræðum, Handritasvið, Safn Árna Magnússonar, Reykjavík.

Zeevaert, Ludger (2014): “Mörkum Njálu! An annotated corpus to analyse and explain grammatical divergences between 14th-century manuscripts of *Njáls saga*.” *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC 14) (Nicoletta Calzolari, Khalid Choukri, Thierry Declerck et al., eds). Paris: ELRA: 981-987. http://www.lrec-conf.org/proceedings/lrec2014/pdf/211_Paper.pdf

Länkade öppna data och den antika världen – två svenska forskningsprojekts deltagande i internationella samarbeten

Johan Åhlfeldt
Centrum för teologi och religionsvetenskap, Lunds universitet
johan.ahlfeldt@ctr.lu.se

Den antika grekiska och romerska världen är till sin karaktär ett internationellt forskningsfält som i Sverige är representerat av ämnen som t.ex. antikens kultur och samhällsliv och religionsvetenskap med subdiscipliner som klassisk arkeologi, numismatik, epigrafik, kyrkohistoria etc. Det är också i just dessa ämnen som det pågår internationella initiativ för att gemensamt komma överens om

auktoritetsdata som syftar till att på ett entydigt sätt referera till antika platser och personer. Det brittiska Pelagios-projektet har samlat ett 30-tal forskningsprojekt och kulturarvsinstitutioner som publicerar data på internet i ett samarbete och har kommit överens om att använda gemensamma och beständiga identitetsbeteckningar för antika platser som förekommer i källmaterial eller som fyndplatser. Detta möjliggör att innehåll i separata databaser kan koordineras med hjälp av gemensamma referenser till plats. Data från Pelagios är öppet tillgänglig och vem som helst kan berika sina forskningsfält med data om källmaterial från bestämda platser från hela nätverket. *Standards for Networking Ancient Prosopographies : Data and Relations in Greco-Roman Names* (SNAP:DRGN) är ett annat brittiskt forskningsprojekt som försöker skapa gemensam virtuell auktoritetsdata för personer i den antika världen, sammanställd från många samverkande forskningsprojekt. Båda projekten kan betecknas som infrastrukturprojekt och använder öppna länkade data för att fullgöra sina syften att skapa en länkad vetenskaplig webb av data om den antika världen.

Två forskningsprojekt vid humanistiska och teologiska fakulteterna vid Lunds universitet, där författaren är verksam, medverkar i dessa samarbeten. Projketet *Digital Atlas of the Roman Empire* (http://dare.ht.lu.se) vid avdelningen för antikens kultur och samhällsliv producerar en fritt tillgänglig digital karta och en databas för antika platser och är en av de kataloger med antika ortsnamn och platser (engelska *gazetteer*) som Pelagios-projeketet erbjuder som auktoritetsdata för att relatera källmaterial i digitala resurser. Den digitala kartan är öppet tillgänglig och används för närvarande av cirka 10 projekt i Pelagios-samarbetet runtom i världen.

Projektet *Early Monasticism and Classical Paideia* vid Centrum för teologi och religionsvetenskap studerar sambandet mellan klassisk bildningstradition och tidigt klosterväsende i det senantika och tidigmedeltida bysantiska riket. Projektet har skapat en textdatabas som kallas ”*Monastica* - ett öppet dynamiskt bibliotek och forskningsverktyg” (http://monastica.ht.lu.se) för senantik bildningslitteratur och annan relaterad monastisk litteratur och utgör en unik samling av texter som bara finns sammanställda här. Genom öppen publicering möjliggör projektet att andra kan ta del av källmaterialet samtidigt som projektet även återanvänder texter och annat källmaterial från andra. Projektet använder gemensamma identitetsbeteckningar från Pelagios och överväger att så småningom använda identitetsbeteckningar för personer från SNAP:DRGN.

Gemensamt för de båda forskningsprojekten vid Lunds universitet är att forskningsdata publiceras enligt principerna för öppna länkade data: 1) de minsta analytiska enheterna i databaserna (t.ex. textsegment och antika platser) kan adresseras med permanenta identitetsbeteckningar, så kallade HTTP URI:er, 2) information om enheterna beskrivs genom RDF-standarder gemensamma för Pelagios-nätverket och andra standarder, och 3) att enheterna länkas till externa enheter med information om samma reella objekt (textsegment i handskrifter och kritiska utgåvor av källtext, antika platser och personer). Vidare tillhandahålls data om enheterna både i maskinläsbar (RDF) och läsbar form (HTML) efter automatisk innehållsförhandling (engelska: *content negotiation*). Detta möjliggör för de båda projekten att vara betydelsefulla delar av en öppet länkad webb av data om den antika världen. Data som sådan speglas inte hos den koordinerande aktörer Pelagios eller SNAP:DRGN, utan endast uppgifter om att data existerar, tillsammans med en kort beskrivning, och att den kan nås från en beständig internetadress. Ansvaret för att uppdatera dessa distribuerade resurser ligger alltså kvar hos den institution som skapat dem. Båda projekten bidrar eller kommer att bidra med nya auktoritetsdata för antika platser och personer som för närvarande saknas så att de gemensamma listorna utökas. Detta möjliggör att andra projekt så småningom kan berika dessa enheter med ytterligare källmaterial.

Pelagios-projektet och dess båda svenska medlemmar kan utgöra en modell för hur vi kan organisera annan tematisk forskning där samverkan genom öppna länkade data eftersträvas och nationella auktoritetsdata inte räcker för att beskriva historiska platser, organisationer och personer i en sammahållen historisk kontext, t.ex. det medeltida Skandinavien eller Stormaktstidens Sverige.

Posters

Digital Editing with Virtual Infrastructures: TextGrid and DARIAH-DE

Mirjam Blümm, Sibylle Söring
Göttingen State and University Library, Germany
soering@sub.uni-goettingen.de

Today, scholarly digital editions represent one of the core application areas of the Digital Humanities. Digital Research Infrastructures and Virtual Research Environments (VREs) enable and support the creation, publication, and long-term archiving of such data. Answering an increasing demand for digital and collective research features in the humanities, the joint projects DARIAH-DE[1] and TextGrid[2], funded by the German Federal Ministry of Education and Research, have for the past years developed in continuous exchange with the research community tools and services which aim at mapping the entire research process.

DARIAH-DE supports digitally-enabled research and teaching in the arts and humanities in the areas of teaching (e.g. training material), research (e.g. best practices, tools for analysis and visualisation), research data (e.g. repository for research data, legal aspects) and technical infrastructure (e.g. collaborative working environments, developer portal).

TextGrid offers a VRE with two main components: The TextGrid Laboratory (TextGridLab)[3] contains a versatile open source software for editing and generating digital sources collaboratively in a protected virtual environment, and allowing for a differentiated user rights management, the TextGrid Repository (TextGridRep)[4] offers an open, XML/TEI-based long-term research archive, in which both the text and image data generated with the TextGridLab, as well as external digital objects, can be published, browsed, explored, analysed, cited, and archived.

A crucial factor of a VRE's success and impact is the use of both technological and semantic standards. The TextGrid architecture supports, amongst common metadata standards, the markup language XML together with the well-established markup data format TEI[5]. They reflect international standards for the sustainable, searchable and reusable mark-up of humanistic sources, especially of digital editions. A differentiated user rights management facilitates collaborative work on a shared project in a non-public environment. Tools, data, and methods can be used mutually, regardless of the operative system, software equipment, or location.

In addition to the tools and services available in the TextGridLab, the TextGridRep provides the user with the possibility to save, publish, and search a variety of digital resources such as XML/TEI encoded texts, images and databases, therefore supporting the creation of Linked Open Data. Thus, both core components of the VRE, the TextGridLab and the TextGridRep, are aligned for an optimal interaction, interlinkeage, and workflow between and with one another.

Beyond the creation of resources, TextGrid ensures the persistent availability of and access to research data as well as optimal interconnectivity, supporting international standards. Collaborative

1 https://de.dariah.eu/
2 https://textgrid.de/en/
3 https://www.textgrid.de/en/registrationdownload/download-and-installation/
4 http://www.textgridrep.de/
5 http://www.tei-c.org/index.xml

research is facilitated by e.g. the annotation of images; further annotation features are currently evaluated.

Thus, TextGrid facilitates the creation of digital editions, from the provision and creation of primary data in XML to published and citable research data; data that can also be made available in an external online portal, or be exported into a print-ready document (PDF).

As of today, DARIAH-DE and TextGrid have about 3.100 registered users and approx. 80 research projects from a broad spectrum of humanistic disciplines ranging from archaeology, art history, philology, cultural studies, medieval studies, jewish and ecclesiastical history to linguistic and musicological studies. Amongst them are single scholars and medium-scale research groups such as "Theodor Fontane's notebooks"[1], as well as large, long-term projects like “IDIOM – Interdisciplinary Dictionary of Classic Mayan”[2].

The paper will show how a complete scholarly workflow can be mapped via a VRE like TextGrid – from collecting and generating primary data through enriching it with metadata and XML/TEI, and finally publishing it in a portal and/or a repository. Technologies allowing for a web based annotation, viewing and publishing of image formats such as Digilib[3] will also be explored.

Additionally, DARIAH-DE and TextGrid Repository will be presented as archives for humanities research data, following sustainable standards and thus allowing for citation, long-term accessibility, further interlinking and scholarly re-use.

Kulturminnena talar – med hjälp av språkteknologiska metoder

Rickard Domeij
Institute for Language and Folklore
Rickard.Domeij@sprakochfolkminnen.se

I Institutet för språk och folkminnens (ISOF) samlingar finns stora mängder material som fram till idag har varit svåra och mödosamma att beforska. Till denna kategori hör det inspelade talmaterialet. Samtidigt finns en tro att språkteknologi och språkdata utgör underutnyttjade resurser som verktyg för att öka dessa materials tillgänglighet.

I denna anda planerar institutet i samarbete med Kungliga Tekniska Högskolan (KTH) och Digisam ett flerårigt projekt som syftar till att använda och utveckla språkteknologiska metoder för åtkomst och förädling av talmaterial.

Inom språktekologin kan man, grovt sett, skilja mellan skriftbaserad språkteknologi – det som oftast åsyftas när termen språkteknologi används – och talteknologi. Talteknologin handlar inte huvudsakligen om text utan om ljud- och videoupptagningar av talat språk och talad interaktion. Samarbetet ska resultera i nya metoder för att utforska kulturarvet i tal som fungerar vägledande för andra kulturminnesinstitutioner (som samordnas av Digisam). Resultaten ska tillgängliggöras genom SWE-CLARIN.

1 https://fontane-nb.dariah.eu/index.html
2 http://www.iae.uni-bonn.de/forschung/forschungsprojekte/laufende-projekte/idiom-dictionary-of-classic-mayan/idiom-english-project-description
3 http://digilib.berlios.de/

Såväl ISOF som KTH utgör parter i SWE-CLARIN, som är en av svenska Vetenskapsrådet finansierad satsning på forskningsinfrastruktur. Det uttalade syftet med SWE-CLARIN är att göra språkbaserade material tillgängliga som primära forskningsdata för de humanistiska och samhällsvetenskapliga (HS) forskningssamfunden med hjälp av de avancerade språk- och talbearbetningsverktyg och språkresurser som har utvecklats under många år inom språkteknisk forskning.

I projektet bidrar KTH med tal- och textteknologisk expertis, medan ISOF bidrar med material, arkivkännedom och forskningskompetens inom HS-området, samt textteknologisk kompetens. Digisam är ett samordningssekretariat för digitalisering, digitalt bevarande och digitalt tillgängliggörande av kulturarvet.

I vår poster kommer vi att närmare beskriva våra projektplaner.

Classical Intertextuality in Late Greek Poetry: a Computational Approach

Erik Henriksson
University of Helsinki
erik.henriksson@helsinki.fi

Contemporary research on late antiquity is gradually transforming our understanding of the cultural and religious developments of this period. Archaeological (e.g. Lavan 2011) and literary studies (e.g. Cameron 2006) have recently shown that in order to appreciate the complex interactions of the late Roman world, we need to deconstruct a long tradition of polemical writing practiced by Christians and other biased groups, and take a fresh look at the primary evidence. As a result, the old conception of a conflict between early Christianity and the classical antiquity is giving way to a more accurate view of a sophisticated dialogue between Christianity, other ancient religions and the classical tradition.

My research project explores these interactions in a corpus of Greek poetry written between c. 300 and 700 CE. With Shorrock (2011), I consider texts with both ”pagan” (denoting all religions other than Judaism and Christianity) and Christian subject matter to spring inherently from the same classical roots, intersecting at several points with each other as well as their classical predecessors. Following Shorrock’s characterization, I draw the distinction between explicitly Christian and non-Christian *modes* of writing, without making any assumptions about the religious beliefs of the authors. My aim is to apply statistical tools to examine the relations between aesthetic ideals and the different modes of writing, using an array of data discussed below. The primary focus is on classical intertextuality, here understood in a broad sense comprising both intentional and unintentional relations between late ancient poetry and its classical models. With this approach, I wish to be able to contribute to the recent discussion (e.g. Agosti 2013) about the relationship between religion and classicism.

Poetic discourse underwent a renaissance in late antiquity, becoming the most significant form of literary communication in the highest spheres of society. This can be illustrated by the huge poetic output of Gregory of Nazianzus (c. 18 000 lines, late 4th century CE), containing a wide range of

subjects from theology to biography. The high status of poetry in late antiquity is also apparent in its official applications: it was typically chosen over prose for such occasions as monumental dedications, panegyrics and Christian congregations. This, together with the fact that a relatively large amounts have survived, makes Greek poetry an interesting object of study in view of the discussion outlined above.

The digital corpus of Greek poetry is provided by the Thesaurus Linguae Graecae project (http://tlg.uci.edu), which has made available most surviving literary Greek texts from Homer until late antiquity and well beyond. However, due to a strict copyright license, no text materials owned by the TLG may be copied into an intermediary database, nor may its own databases be queried via a remote script for a large-scale analysis. One of the methodological advances I propose in this project is a new way to leverage TLG's resources. In my approach, data collection happens client-side through TLG's web interface and includes the collection of morphological analyses, lemmas, word/sentence boundaries as well as metrical and phonetic data. These data, in virtue of excluding the original text, may legally be saved to a database to be used in the actual analysis, opening up new possibilities for a quantitative study of TLG's materials.

Methodologically, my project draws from recent developments in using quantitative models to automatically detect intertextuality (e.g. Coffee et al. 2013). Most existing methods have a full-text corpus at hand, whereas this project (for the reasons discussed above) is constrained to a relatively limited collection of data. In many cases, however, an algorithm can be modified to suit different needs. For example, Forstall et al. (2011) have proposed a "functional n-gram" feature to detect influence using character-level n-grams; but as the authors point out, the same approach can be used with other feature types as well. My aim is to employ a similar method with the data that can be extracted from a TLG text, such as metrical patterns, word boundaries and phonetic classes.

This poster will present new tools to help collecting a number of data from a poetic text provided by the TLG. These include an algorithm which computes the most probable metrical reading of a given line, an integrated morphological analysis tool to speed up the collection of lemmata, as well a front-end to manage the collected datasets in a Text Encoding Initiative (TEI)-compliant XML format.

References

Agosti, G. (2013). 'Classicism, *paideia*, Religion'. *The Strange Death of Pagan Rome.* Rita Lizzi Testa (ed.). Turnhout: Brepols. https://www.academia.edu/9213407.

Cameron, A. (2006). 'Poetry and Literary Culture in Late Antiquity'. *Approaching Late Antiquity: The Transformation from Early to Late Empire.* Simon Swain, Mark Edwards (eds.). Oxford: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199297375.001.0001.

Coffee, N. et al. (2013). 'The Tesserae Project: Intertextual Analysis of Latin Poetry'. *Literary and Linguistic Computing*, **28** (2): 221–228. http://llc.oxfordjournals.org/content/28/2/221.

Forstall, C. W., Jacobson, S. L., Scheirer, W. J. (2011). 'Evidence of intertextuality: investigating Paul the Deacon's *Angustae Vitae*'. *Literary and Linguistic Computing*, **26** (3): 285-296. http://llc.oxfordjournals.org/content/26/3/285.

Lavan, L. (2011). 'Introduction'. *The Archaeology of Late Antique "paganism"*. Luke Lavan, Michael Mulryan (eds.). Leiden: Brill.

Shorrock, R. (2011). *The Myth of Paganism.* London: Bloomsbury.

Comparative palaeography of cuneiform: a new digital humanities method and approach

Zenobia Homan
School of Oriental and African Studies, London
603466@soas.ac.uk

Traditionally, cuneiform palaeography has been accomplished via meticulous hand-illustrating of tablets based on years of experience. Digital methods are still novel. However, my statistical approach to the subject has taken the form of a unique python-programmed database of sign-instances extracted from photographs. It is possible to organise sign-groups via dragging and dropping images, and the database also counts the number of signs per tablet. The new program allows for the establishment of visual parameters of variation, and the calculation of percentages (such as the occurrence of different sign-forms). Using these data it is possible to create graphs illustrating sign-form frequency as well as combined frequency (co-occurring features), and in turn it is possible to comment on scribal culture and individual scribal hands. The purpose of the database is to offer a more objective approach to comparing archives than previously possible. My research specifically aims to determine the features of Mittani (and 'Assyro-Mittanian') cuneiform, compared to Middle Assyrian, Babylonian and Nuzi. Naturally there are many underlying problems which must be addressed. The terminology of signs, sign-forms and sign-variation still lacks firm footing, and the parameters between them are open to debate. In addition, this database was developed for the study of cuneiform specifically: it would be interesting to explore how it may be used with different script-types, to work towards more universally applicable digital methods in palaeography.

Att hantera felegrafen: Textanalys av smutsiga tidningar

Johan Jarlbrink, Roger Mähler & Pelle Snickars
HUMlab
Umeå university
johan.jarlbrink@umu.se, roger.mahler@umu.se, pelle.snickars@umu.se

Digitala metoder för textanalys, såsom samförekomstanalys och topic modeling, utgår ofta från bok-och tidskriftspublicerade texter som digitaliserats, eller material som är digitalt fött. Texterna från böcker och vetenskapliga tidskrifter är ofta ”rena” i den meningen att texterna löpt i enkla eller dubbla spalter över sidorna, vilket underlättar digitalisering. Texterna är relativt enhetliga vad gäller stilsort, och det är sällan flera olika texter på samma sida. Skönlitterära texter har därtill ofta granskats innan de gjorts digitalt tillgängliga. Detta innebär att OCR-felen är begränsade och att olika texter, exempelvis artiklarna i en tidskrift, på ett enkelt sätt kan skiljas från varandra. Digitala metoder för textanalys, som vanligen bygger på att OCR-felen är få och att textsegmenteringen är

korrekt, fungerar utmärkt för sådant material. De texter som har digitalt ursprung är förstås ännu renare i dessa avseenden.

Alla texter som nu finns digitalt tillgängliga är emellertid långt ifrån lika rena. Hur kan man arbeta med betydligt smutsigare texter? I denna presentation beskriver vi hur vi arbetat med svenskt digitaliserat tidningsmaterial från 1800-talet. Nationalbibliotek runt om i världen håller just nu på att digitalisera sina samlingar av äldre tidningspress. Mycket har redan tillgängliggjorts; mer är att vänta. I Australien finns exempelvis nära 20 miljoner tidningssidor sökbara i databasen ”Trove”. I USA finns omkring 10 miljoner sidor tillgängliga via Library of Congress. I Norden erbjuder Finlands Nationalbibliotek tre miljoner sidor, och Danmark 10 miljoner. I Sverige finns omkring 10 miljoner sidor sökbara, men de flesta är från senare tid. Endast fyra tidningstitlar och mindre än 200 000 sidor 1800-talspress har digitaliserats. På grund av en strängt tolkad upphovsrättslagstiftning är det endast detta äldre material fritt sökbart online i Sverige.

Digitaliseringen innebär nya möjligheter att söka och hitta – men felkällorna är många. Detta märker man inte minst om man använder digitala metoder för att söka efter mönster i textmassorna utan att gå via bibliotekets gränssnitt. 1800-talets tidningssida är mycket olik den romansida som många av fältets föregångare har arbetat med. En enskild tidningssida kan i vissa fall innehålla hundratalet korta notiser, telegram och radannonser. Dessa texter löper i vissa fall i så många som sju täta spalter. De redaktionella texterna inleds sällan med rubrik, och ibland är det inte ens en tunn linje som markerar var en text börjar och slutar. I annonsmaterialet finns också stor grafisk variation. Flera olika stilsorter och storlekar kan förekomma i en och samma mening. Fraktur och antikva kan blandas om vartannat. Dessa kaotiska textkollage ställer OCR-teknik och autosegmentering inför betydande prövningar. De fellästa orden är många, och kortare texter har i regel systematiskt klumpats ihop med varandra, medan längre texter har hackats upp i mindre delar. De textblock som autosegmenteringen identifierat som avgränsade enheter kan exempelvis bestå av en halv ledarartikel om skattepolitik sammanförd med annonsen för nya korsetter. De autogenererade ”topics” som baseras på dylika textblock kommer därför att se mycket märkliga ut. Också enklare varianter av samförekomstanalys ger konstiga resultat: med ett textblock som ovan kommer det att se ut som att skattefrågan hänger ihop med korsetter.

I en studie om den elektriska telegrafen i *Aftonbladet* 1830–1863 har vi försökt hitta metoder för att hantera dessa problem. Vi har haft direkt tillgång till de textblocksindelade textfilerna. För att hitta relevanta texter har vi sökt med ett Levenshteinavstånd som tillåter ett visst antal OCR-fel i de eftersökta nyckelorden, ”elektrisk” respektive ”telegraf”. Efter en manuell granskning av ordträffarna har listorna med de tusentals varianterna av nyckelorden använts som sökord för att hitta textblock där båda orden förekommer. Därefter har textblocken granskats manuellt för att sortera ut vilka texter inom varje textblock som verkligen handlar om den elektriska telegrafen. Dessa texter har använts för en samförekomstanalys, i syfte att fånga centrala begrepp och aktörer i artiklarna om den nya teknikens etablering. Vi har använt oss av flytande textfönster för att beräkna samförekomst. Vi kan då undvika de felaktigheter som finns med den ursprungliga textblocksindelningen. Resultatet är inte problemfritt, men tillvägagångssättet gör att en del av textmaterialets brister kan hanteras. Syftet med denna presentation är alltså att belysa textmaterialets kvalitetsbrister – samt att presentera metoder som gör texterna möjliga att analysera digitalt.

Johan Jarlbrink är docent i mediehistoria och lektor i medie- och kommunikationsvetenskap på Institutionen för kultur- och medievetenskaper, Umeå universitet, sam affilierad forskare vid HUMlab. Roger Mähler är utvecklare vid HUMlab. Pelle Snickars är professor i medie- och kommunikationsvetenskap med inriktning mot digital humaniora på Institutionen för kultur- och

medievetenskaper och affilierad forskare vid HUMlab.

Digital Humanities in the field of Linguistics and Language Research

Janne Bondi Johannessen, Kristin Hagen, Michał Kosek, Anders Nøklestad, Joel Priestley
Text Laboratory, Iln, University of Oslo
j.b.johannessen@iln.uio.no, kristin.hagen@iln.uio.no, michal.kosek@iln.uio.no, anders.noklestad@iln.uio.no, joeljp@gmail.com

This paper looks at the requirements and possibilities of tools for corpus linguistic research and text analysis, presenting Glossa as one case for meeting such challenges.

Glossa was originally developed in 2005 at the Text Laboratory, University of Oslo, to provide a search interface for parallel corpora. Although powerful, the tools available at the time were most relevant for the tech-savvy linguist; the intention behind Glossa was to facilitate access to the broader community. This was achieved by creating a web interface consisting of familiar browser components from which, behind the scenes, appropriate queries for both databases and corpus search systems could be built. Over the last decade, Glossa has developed in response to an influx of requests for new features and applications. Most fundamental to this evolution was a thorough re-implementation under the Clarin umbrella, which began in 2014 and is due to be completed by spring 2016.

The rewrite was motivated by the need to maintain a flexible, modular system, suitable for continual adaptation to new demands and in particular to satisfy the Clarin requirement of enabling federated content search. Another important consideration was one of distribution. The earliest version was viewed more as an in-house facility; a single instance capability for providing access to the various corpora hosted by the Text Laboratory. Soon, however, requests were received from organizations not affiliated with the University of Oslo who wished to host their own corpora using the same system. Owing to this rapid and unforeseen development, the original codebase for Glossa became unwieldy, making installation challenging for those unfamiliar with its components. Additionally, personal computers were not a target for Glossa installation because it was developed as a web server application. It became apparent that such functionality would be useful, particularly following experience gained from collaboration with Addis Ababa University, which showed that a personal installation could be an invaluable tool for field work when internet access is unreliable or non-existent.

Before commencing on the rewrite, it was important to reconsider the alternatives. Since Glossa's inauguration, the landscape had inevitably changed and at least one other comparable system had arrived on the scene. Språkbanken, at the University of Gothenburg, had developed their own corpus search tool, called Korp. Both conceptually and functionally, there is much overlap between Korp and Glossa. However, experience gained through years of adapting Glossa to ever new areas of corpus linguistics, tipped the scales in favour of rebuilding our system from scratch. There was a desire to reap the benefits of lessons learned, as well as keep full control over the codebase and hone in-house skills. Glossa had, after all, been a success. In particular, the innate focus on spoken corpora that Glossa has, weighed heavily.

Glossa has attracted interest and requests from a continually wider section of the linguistically interested, from a range of disciplines including musicology, dialectology, translation research,

lexicography, medical science, and second language acquisition. Each new field of application brings with it new requirements, enriching Glossa with new features such as geographical maps, acoustic waveforms and spectrograms, statistics and multimedia capability. These features will be demonstrated, along with a selection of the corpora that prompted their development, including:

- Nordic Dialect Corpus - a spoken corpus of roughly 2.5 million words that utilises both audio/video, waveforms/spectrograms and maps
- Corpus of music reviews - 271 Norwegian language music reviews
- The Mandarin Audio Idiolect Dictionary (MAID)
- The RuN Corpus - a Russo-Norwegian parallel corpus of 1.2 million words for each language
- NoWaC - a web corpus of the Norwegian *.no domain, comprising some 700 million words

At its simplest, Glossa provides a means of performing text searches in corpora. Depending on the nature of the corpus investigated, Glossa will provide an array of search criteria, giving access to multiple layers of transcription, including morphology, orthography, phonetics, and aligned translation. Where applicable, constraints may be applied using metadata associated with the transcription, such as text genre, location, date and demographics. Similarly, the mode in which results are delivered may be varied in accordance with the complexity of the material, from simple KWIC concordances to multi-layered renditions of the original text. Frequency lists and collocations can be quickly distilled from the result set, which can be further edited, sorted and processed with user annotations. Where audio/video files are provided, fine-grained acoustic analyses can be performed, and video with autocue style subtitling can be displayed. Geographical metadata can be rendered on a built-in map feature, aiding the detection of isoglosses. The evolution of Glossa is ongoing; areas of development include graphic visualizations and the use of maps as part of the search interface.

The current version of Glossa is implemented with the Lisp dialect Clojure/ClojureScript. Clojure runs on the Java Virtual Machine while the ClojureScript compiler targets JavaScript. While the Text Laboratory has chosen the IMS Open Corpus Workbench and MySQL to power Glossa, the system is inherently agnostic in terms of this backend.

From Handwritten Slips to a Digital Dictionary Corpus

Ellert Johannsson, Simonetta Battista
University of Copenhagen
nkv950@hum.ku.dk, sb@hum.ku.dk

A Dictionary of Old Norse Prose (ONP) is a dictionary project based at the Department of Nordic Research at the University of Copenhagen. ONP records the vocabulary of prose writing in Old Norse, as transmitted in Norwegian and Icelandic manuscripts, the earliest of which date from the middle of the 12th century.

The project started in 1939 and was originally intended to supplement the older lexicographic works of Fritzner (1886, 1891, 1896) and Cleasby/Vigfusson (1874), but it quickly evolved into an independent lexicographic project. For some decades work on the dictionary consisted in selected excerption of texts covering all Old Norse prose genres, with special attention to oldest attestations, linguistic structures and variation, as well as culturally significant vocabulary. The result was an

extensive archive of around 750.000 citations on handwritten slips, where the original orthography of the source is rendered in detail. This citation corpus constituted the basis for subsequent editorial work.

An originally planned publication of a 13 volume dictionary commenced in 1989 but was put on hold in 2004 after a volume of indices (ONP:Registre) and three volumes of dictionary entries covering words from a-em (ONP1, ONP2, ONP3) had been published.
In 2005 the decision was made to change ONP into a digital publication, which was meant to include both the material from the published volumes as well as unpublished citations. The challenge was to digitalize this material and make it accessible and useful for the users of the dictionary. ONP Online was launched in 2010.

This poster presentation gives an overview of the process of converting the analogue archive of handwritten citation slips into a digital citation corpus, which now forms the core of the online dictionary. This process included the following steps, some of which are still ongoing:

- preparing a digital list of the dictionary headwords
- scanning of citation slips under the appropriate headword
- updating references to accommodate newer text editions
- scanning of text editions
- data entry of citations

As a result the users of ONP Online have access to all the available citations compiled under each headword, where each citation is linked to a scanned page from a scholarly text edition.
In the presentation we also illustrate how ONP's material can be searched and used in various ways and present further possibilities in developing and exploiting the digital citation corpus. We discuss the underlying database, which has been especially tailored for this dictionary project, and explore its potential usage in order to provide more personalized search capabilities. Finally, we mention some future possibilities of expanding the digital citation corpus, e.g. by incorporating recent digital editions of Old Norse texts (cf. MENOTA project) into ONP Online.

Bibliography

Cleasby, Richard & Gudbrand Vigfusson (1874). *An Icelandic-English Dictionary*, Oxford: Clarendon Press.
Fritzner, Johan (1886, 1891, 1896). *Ordbog over Det gamle norske Sprog* 1-3, rev. ed., Kristiania: Den norske forlagsforening.
MENOTA. *Medieval Nordic Text Archive*. Accessed at: www.menota.org [17/11/2015]
ONP Online. Accessed at: www.onp.ku.dk [13/11/2015].
ONP:Registre = Ordbog over det norrøne prosasprog: Registre / A Dictionary of Old Norse Prose: Indices. (1989). Copenhagen: The Arnamagnæan Commission.
ONP1 = Helle Degnbol, Bent Chr. Jacobsen, Eva Rode, Christopher Sanders & Þorbjörg Helgadóttir (1995). *Ordbog over det norrøne prosasprog / A Dictionary of Old Norse Prose* 1: a-bam. Copenhagen: The Arnamagnæan Commission.
ONP2 = James E. Knirk, Helle Degnbol, Bent Chr. Jacobsen, Eva Rode, Christopher Sanders & Þorbjörg Helgadóttir (2000). *Ordbog over det norrøne prosasprog / A Dictionary of Old Norse Prose* 2: ban-da. Copenhagen: The Arnamagnæan Commission.
ONP3 = Helle Degnbol, Bent Chr. Jacobsen, James E. Knirk, Eva Rode, Christopher Sanders & Þorbjörg Helgadóttir (2004). *Ordbog over det norrøne prosasprog / A Dictionary of Old Norse Prose* 3: de-em. Copenhagen: The Arnamagnæan Commission.

Hassle-free creation of metadata and conversion of text material to TEI format

Bart Jongejan, Lene Offersgaard
University of Copenhagen
bartj@hum.ku.dk

Introduction

Since 2012 researchers have had the opportunity to deposit resources in the CLARIN-DK repository (Offersgaard et.al. 2013). The deposit service accepts resources and accompanying metadata in a limited number of formats. About forty-thousand texts have been deposited, but to date, almost no individual researchers have made the conversions of their data and metadata that are required before depositing can be undertaken. Instead, the preparation of all these resources for depositing has been done by the CLARIN-DK team. Dialogues with researchers in humanities have revealed that providing data in the required formats and defining metadata is too hard for many of them.

Much of the textual data that researchers might want to share with others is not in one of the formats that CLARIN-DK accepts as depositing format. Converting a resource to the required format involves a number of steps, requiring a mix of skills that few have. Some resources require scholarly diligence, for example if a text is manually annotated in a free style, or if a text contains elements that the researcher does not want or is not allowed to share. After this manual step, one or more steps follow that require technical prowess that not all scholars have.

The CLARIN-DK team has validated the content of the deposited material and made the observation (Hansen et.al. 2014) that it is much more difficult to supply data with metadata in a consistent way than expected.

Web service for easy conversion of text to TEI format and creation of metadata

To reach out to many scholars with a solution to these problems, we have extended the information available online, created a helpdesk and made a web service that makes providing data and metadata of acceptable quality much easier. The web service transforms uploaded text to the TEIP5 format and asks the user for an amount of metadata that is easy to provide and at the same time informative enough for other researchers when they search and browse through the repository.

The CLARIN-DK text preparation service can handle uploaded text in a still growing variety of formats: plain text, office documents, web documents, PDF documents, and images. The CLARIN-DK conversion service uses third party open source software to convert the input to plain text or to RTF, from where other tools, developed by the CLARIN-DK team, take over and complete the creation of TEI-output. If the need arises, the CLARIN-DK staff can implement specialized transformation tools for text data that require an idiosyncratic treatment and add those tools to the existing set of integrated tools. Once these tools are registered (URL, input and output specifications) they become part of the CLARIN-DK ecosystem and are automatically incorporated in workflows when needed. (Jongejan 2013)

After uploading a text resource for conversion, the user is confronted with a single page of metadata entry fields. Each field is accompanied by a help text in Danish or English. The web form does not require any XML and XSLT skills or knowledge of schema validation. The web form asks for all

obligatory metadata, which are fewer than before. In addition, there is a selection of optional metadata fields. Each field is validated and the user is asked to correct any errors that are discovered.

After completing the web form the conversion of the uploaded text starts. The user downloads the results to her computer and inspects the end product, which is a TEIP5 file containing a metadata header and a body with text. Also the intermediary results for each conversion step are available to the user in easily accessible formats: plain text, RTF, or HTML. If needed, corrections can be made in the intermediary files, which then can be uploaded to the conversion web service.

Future

We will implement a facility for previewing the output, so the researcher can decide to deposit the created resource or to correct errors first. This preview will show the text itself and the metadata.

Initially we only address the preparation of text resources, but after gathering some experiences we will evaluate this service and extend it to other resource types.

We will implement a facility that allows experts to extend the metadata with more details.

Conclusion

With a web-based text preparation service we reach out to researchers who hitherto have been unable to share their textual data due to limited resources, while ensuring that textual data and their metadata are made available in a uniform fashion.

References

Hansen, D. H., Offersgaard, L., & Olsen, S. (2014). Using TEI, CMDI and ISOcat in CLARIN-DK. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation: LREC 2014*. (pp. 613 - 618). Reykjavik, Iceland: European language resources distribution agency.

Jongejan, B. (2013). Workflow Management in CLARIN-DK. In *Proceedings of the workshop on Nordic language research infrastructure at NODALIDA 2013*. (Vol. 089, pp. 11-20). Linköpings Universitet: Linköping University Electronic Press. (NEALT (Northern European Association of Language Technology) Proceedings Series, Vol. 20).

Offersgaard, L., Jongejan, B., & Hansen, D. H. (2013). CLARIN-DK – status and challenges. *Proceedings of the workshop on Nordic language research infrastructure at NODALIDA 2013*. (Vol. 089, pp. 21-32). Linköping University Electronic Press. (NEALT (Northern European Association of Language Technology) Proceedings Series; Vol. 20).

Scribes and Late Latin Charters: a treebank-based study of the language competence of scribes

Timo Korkiakangas
University of Oslo
timo.t.korkiakangas@gmail.com

In my poster I present an incipient DH-oriented project on the Latin of Early Medieval documentary

texts (charters). I aim to use treebanking to track the influence of the production environment as a factor of language change in historical text corpora. The technical objectives include finding optimal language competence metrics (normal corpus-linguistic methods and novel computational measures of spelling variation and syntactic complexity) as well as creating a network model. These will be utilized to estimate the scribes' effective language skills, language attitudes, and stylistic as well as morphosyntactic preferences by mirroring the writing performance with the geographical location, time of writing, and social context (whenever known).

This far, the data consist of 519 Italian (Tuscan) private documents from between AD 714–869, which together form the 200,000-word Late Latin Charter Treebank (LLCT). A parser will be trained on the existing annotations to upgrade LLCT to LLCT2 by ingesting 135,000 additional words from other charters. The treebank contains lemmatic, morphological, syntactic (dependency grammar), and light semantic annotation layers plus some spatio-temporal metadata. I have followed the *Guidelines for the Syntactic Annotation of Latin Treebanks* (Perseus Latin Dependency Treebank, LDT) with several additions concerning the nonstandard morphology of the charters (Korkiakangas & Passarotti 2011). The treebank is in the Ancient Language Dependency Treebank format (ALDT). Charters are privileged material for examining the spoken/written interface of Early Medieval Latin, in which the distance between the spoken and written codes had grown considerable by the Late Antiquity. The LLCT charters have precise dating and location metadata and they survive as originals, i.e. with no transmission history.

Previous studies show that the charter Latin reflects some developments of the spoken language. My hypothesis is that the emergence of innovative spoken-language-born linguistic forms or structures (and, respectively, the retention of conservative forms or structures) in written code correlates positively with the spelling competence of the scribe and with the level of syntactic complexity he used to produce: the better a scribe spelled, the fewer spoken-language related innovations and the longer dependencies are expected in his output. I will examine at least the following innovative or conservative phenomena: the use of the Classical Latin case repertoire (high vs. low retention of the system, substitution by PPs, morphosyntactic alignment), subordinate constructions (ACI vs. clausal), number and quality of pronoun determinants of NPs, and word order (SOV vs. SVO). All the phenomena can be retrieved from the morphosyntactic and semantic markup of LLCT2 with corpus-linguistic methods. I seek to answer the following research questions: A) Are there statistical dependences between the spelling level variable and the other language competence variables? B) Are there statistical dependences between the language competence variables and the contextual variables? C) What kind of upper-level patterns does the network visualization reveal between the document and actor (i.e. scribe) nodes and the language competence metrics?

Technically, the project comprises of three major work packages: 1) The spelling competence metric will be defined by measuring the edit distance between non-standard and normalized word forms (in proportion to the word count of each charter text). The normalization of each lemma/morph-tag pair of LLCT2 will be realized by applying an inflected-word-form generator based on the Morpheus code (Perseus Project). 2) The syntactic complexity metric will be obtained by measuring dependency distances, i.e. the word count between a dependency and its parent word in each charter text. Average dependency distance gives a rough measure of a text's syntactic complexity and (arguably) its difficulty. 3) The primary ambition of the project is "linguistic" network modelling, i.e. visualization of the connections between linguistic phenomena and the mentioned contextual (socio-historical) variables. The spelling, dependency distance, and morphosyntactic metrics will be presented as interval variables and displayed as graduated nodes.

Bamman, D. & al. 2007. *Guidelines for the Syntactic Annotation of Latin Treebanks* (v. 1.3). http://nlp.perseus.tufts.edu/syntax/treebank/1.3/docs/guidelines.pdf

Bartoli Langeli, A. 2006. *Notai: scrivere documenti nell'Italia medievale*. Roma.

Korkiakangas, T. & Passarotti, M. 2011. 'Challenges in Annotating Medieval Latin Charters', in *Journal of Language Technology and Computational Linguistics* (JLCL) 26:2, 103–114. http://www.jlcl.org/ 2011_Heft2/16.pdf

Korkiakangas, T. & Lassila, M. 2013. 'Abbreviations, fragmentary words, formulaic language: treebanking medieval charter material', in F. Mambrini & al. (eds.) *ACRH-3 Proceedings*, 61–72.

Liu, H. 2008. 'Dependency distance as a metric of language comprehension difficulty', in *Journal of Cognitive Science*, 9:2, 159–191. http://www.researchgate.net/publication/273459859_Dependency_Distance_as_a_Metric_of_Language_Comprehension_Difficulty

Passarotti, M. 2014. 'The Importance of Being sum. Network Analysis of a Latin Dependency Treebank', in *CLiC-it 2014 Proceedings*, 291–295. http://clic.humnet.unipi.it/proceedings/vol1/CLICIT2014156.pdf

Setting up a Digital Humanities Centre in Romania. Problems and solutions

Corina Elena Moldovan, Voica Puscasiu
Babes-Bolyai University
voicapuscasiu@gmail.com

This paper is a commentary on the manner in which a DH Centre is established in an Eastern-European country and the challenges that exist within the traditional structures and mentalities at different levels - the academia, the policy-makers and the society in the larger sense. The poster will focus on the strategies that DigiHUBB has, at different levels – education, research and innovation, collaboration with the IT industry. It will also point out similarities between other post-revolutionary experiences and learn if there could be a kind of specificity that could give our centre a uniqueness that will make it relevant internationally. The argument in this paper resonates strongly with the thematic implications addressed in this conference – the "social, institutional and multicultural aspects of digital humanities".

GROUP INTERACTION AND SYNCHRONISATION TO MUSICAL PASSAGES OF ELECTRONIC DANCE MUSIC

Ragnhild Torvanger Solberg,
University of Agder
ragnhild.t.solberg@uia.no

Alexander Refsum Jensenius

University of Oslo
a.r.jensenius@imv.uio.no

How do dancing crowds engage with Electronic Dance Music (EDM) in a club-like environment, and how can such experiences be measured? This paper reports on an empirical study of dancers' experience of particular musical passages, more specifically their embodied engagement associated with the *break routine* of EDM. This routine consists of three musical passages: (1) breakdown, (2) build-up and (3) drop, which leads to shifting intensity and energy in the music due to large structural and textural changes.

29 participants (15 females and 14 males, M: 22.2 years, SD: 6.5) distributed onto three different groups danced to a short DJ mix in a club-like atmosphere. This atmosphere consisted of a dance floor with light effects and a good sound system created in a motion capture lab. The groups' body movements were recorded and measured using an infrared marker-based motion capture system and accelerometers. The subjects filled out a questionnaire after the dancing session, with information on their desire to dance during the experiment and its association to specific musical characteristics.

The results of the analyses show correspondences between the groups' level of movement and the break routine. Especially after the drop moment, the groups' movements grew larger, more pronounced and more in synchronisation with the beat of the music. Interestingly, the participants reported about the same musical passages, such as the build-up and drop, as causing an arousal in their desire to dance. This indicates that these passages "energize" the groups and their dancing.

One of the key features of the break routine is to temporally remove and then reintroduce the rhythmical patterns of the music, especially the bass drum. Thus our results further suggest that the presence of the beat keep a group more synchronised than when they are lacking their rhythmic frame of reference. In conclusion, these findings show how infrared motion capture and accelerometers are fruitful ways to investigate musical experiences of groups in a controlled yet ecological valid setting. Through these systems we can obtain detailed data about both the quantity and quality of a group's sound-accompanying body movements.

Digitizing And Opening Access To Six Icelandic Dictionaries

Steinþór Steingrímsson
The Árni Magnússon Institute for Icelandic Studies
steinst@hi.is

The paper describes a project in progress. Six Icelandic dictionaries are being digitized and will be made accessible on the web. A web platform is being built to allow searching and browsing the dictionaries. Some of the dictionaries will also be made available to download in a machine readable format.

The six dictionaries to be made accessible online are:

1) The *Icelandic etymological dictionary* compiled by Ásgeir Blöndal Magnússon in the late 20th century and originally published in 1989.
2) The *Icelandic-Danish dictionary* by Sigfús Blöndal, published in 1920, with supplement published in 1963.
3) *Dictionary by Gunnlaugur Oddsson*. A Danish-Icelandic dictionary first published in 1819. Second edition edited by Jón Hilmar Jónsson and published in 1991.
4) *Icelandic-Latin-Danish dictionary by Björn Halldórsson*. Compiled late in the 18th century and first published in 1814. Second edition edited by Jón Aðalsteinn Jónsson and published in 1992.
5) *Nucleus Latinitatis*. A Latin-Icelandic dictionary translated by Jón Árnason from the Icelandic-Danish dictionary by Hans Gram, published early in the 18th century.
6) *Lexicon Islandicum*. An Icelandic dictionary with definitions in Latin. Compiled by Guðmundur Andrésson between 1650 and 1654 and first published in 1683. Second edition was edited by Jakob Benediktsson and Gunnlaugur Ingólfsson and published in 1999.

The motivation for making these dictionaries accessible online is to make it easier for researchers, students, lexicographers and others around the world, working with the Icelandic language or interested in the history of the language to use these out-of-print resources.

At the beginning of the project the dictionaries were at different stages of digitization. All but the Icelandic Danish dictionary have been printed in the last 25 years and were all layed out using LaTeX. The LaTeX documents still exist and the markup is quite regular, so relatively little preprocessing is necessary before splitting the dictionary into articles, which are inserted into a database for search and lookup from a web page. The Icelandic-Danish dictionary on the other hand is only available in print and has never been digitized in any way. Thus we had all the 1100+212 (original+supplement) pages photographed. Then we use ABYYY FineReader 12 to OCR-read the pages. The resulting text is then proofread, the articles separated and markup added before it is published online. We estimate the text processing to be approximately three months work.

On the web platform, the user can search or browse one dictionary or all the dictionaries at the same time. Pages for each of the dictionaries will provide information and explanations for the corresponding work, a short introduction to the work, explanations for abbreviations and other information that helps the user use the dictionary in question.

Furthermore five of the dictionaries (2-6 in the list above) will be made available to download in a

machine readable format.

We plan to make the dictionaries available one by one throughout 2016. The first one, The Etymological dictionary (no. 1 in the list above), will be available by the end of January. Dictionaries 3-6 in the list above are planned to be made available by the end of April 2016 and the Icelandic-Danish dictionary by the end of October 2016. The platform will be available on ordabaekur.arnastofnun.is.

Digital Humanities in Swedish research: a bibliographic review

Anna Svensson
Gothenburg University Library, Humanities Library
Anna.Svensson@ub.gu.se

This poster presents a bibliographic review that seeks to investigate the extent to which Swedish research is represented in the Digital Humanities (DH) through the analysis of: 1) the content of a selection of bibliographic databases; and 2) the content in some of the journals highlighted by DH organisations.
The databases used are International Bibliography of the Social Sciences (IBSS), Linguistics and Language Behavior Abstracts (LLBA), Modern Language Association (MLA), Scopus, SwePub and Web of Science. Searches were made for '"digital humanities" OR "humanities computing"[1] OR "digital humaniora"' for the publication types journal article or review article in the fields for document title, abstract or keywords. The searches generated a total of 678 references, 26 of which were written by researchers with Swedish affiliation; excluding SwePub there were 7 references. Taking into consideration all publication types in SwePub, there were 65 references retrieved. For comparison a broader search[2] was made in MLA where analysis of the keywords revealed that "computer-assisted research" and "digital technology" were more common as keywords in the 67 references that related to Swedish research.
From the webpages of Alliance of Digital Humanities Organizations (ADHO), European Association for Digital Humanities (EADH) and Digital Humanities in the Nordic countries (DHN), the following journals where selected and revised for Swedish research[3]:
Digital Humanities Quarterly, *DHQ* 2007- (ADHO, EADH, DHN); *Literary and Linguistic Computing*, *LLC* 1986-2014 / *Digital Scholarship in the Humanities*, DSH 2015- (ADHO, EADH, DHN); *Digital Studies/Le champ numérique*, *DS* 2009- (ADHO); *Computers and the Humanities,* 1966-2004 / *Language Resources and Evaluation*, 2005- (EADH); *Journal of the Text Encoding Initiative*, 2011- (ADHO) and *Human IT*, 1997- (DHN).
There were 2658 articles in the journals after excluding book reviews, editorials and introductions, 181 were written by researchers with Swedish affiliation and 173 of these were not found in the

1To include "humanities computing" in the database search might bias the result towards certain aspects of the digital humanities, but it gives a view of the shift in time of the use of the terms.

2(ti(comput* OR digital* OR "text encoding" OR "text mining" OR "data mining" OR n-gram* OR "topic modeling" OR gis OR "Geographic Information System") OR mlasu(comput* OR digital* OR "text encoding" OR "text mining" OR "data mining" OR n-gram* OR "topic modeling" OR gis OR "Geographic Information System") OR abs(comput* OR digital* OR "text encoding" OR "text mining" OR "data mining" OR n-gram* OR "topic modeling" OR gis OR "Geographic Information System")) AND (sweden OR swedish) for Journal Articles 1968-2015

3http://adho.org/Publications ; http://eadh.org/publications/all ; http://dig-hum-nord.eu/?page_id=87&lang=sv as of November 2015. The three journals first mentioned are also highlighted by the European Science foundation in their report on research infrastructure in humanities (Moulin 2011, 28).

database search for “digital humanities” or “humanities computing”. If we exclude *Human IT*, there were 29 articles.
The bibliographic corpus from the databases reflects the use of the term DH in the metadata, which indicates that as a search term it has complications. In a similar search in Web of Science, Leydesdorff y Salah (2010) pointed out that the relatively few hits found where due to fact that the database contains mostly formal literature. We suggest that it’s also due to the relatively amorphous character of the concept itself, reflected in the abundant literature on defining the field, in self definitions projects like “Day of DH” and, as Patrik Svensson (2010) suggests, that even researchers doing DH might not identify with the concept. Other factors could be practices of indexing such as the aim for specificity or inconsistencies in assigning keywords on methods or tools used for the study indexed.
The poster visualizes some characteristics of the corpus including timeline, topics, journals and affiliations. The references related to Swedish research are published in a public Zotero group.
https://www.zotero.org/groups/digital_humanities_in_swedish_research_dhn_2016
The review is based on an updated bibliographic corpus used for an article on the representation of Latin American Studies within the DH (Svensson 2014).
References
Leydesdorff, Loet, and Alkim Almila Akdag Salah. 2010. "Maps on the basis of the Arts & Humanities Citation Index: The journals Leonardo and Art Journal versus “digital humanities”as a topic." *Journal of the American Society for Information Science and Technology* 61 (4):787-801. doi: 10.1002/asi.21303
Moulin, Claudine, ed. 2011. *Research infrastructures in the Digital Humanities*. Strasbourg: European Science Foundation. http://www.esf.org/hosting-experts/scientific-review-groups/humanities-hum/strategic-activities/research-infrastructures-in-the-humanities.html
Svensson, Anna. 2014. “El término Humanidades Digitales y los Estudios Latinoamericanos: una revisión bibliográfica”. *Anuario Americanista Europeo* 12: Sección Documentación 1-28. http://www.red-redial.net/revista/anuario-americanista-europeo/article/view/281
Svensson, Patrik. 2010. "The landscape of Digital Humanities." *Digital Humanities Quarterly* 4 (1). http://www.digitalhumanities.org/dhq/vol/4/1/000080/000080.html

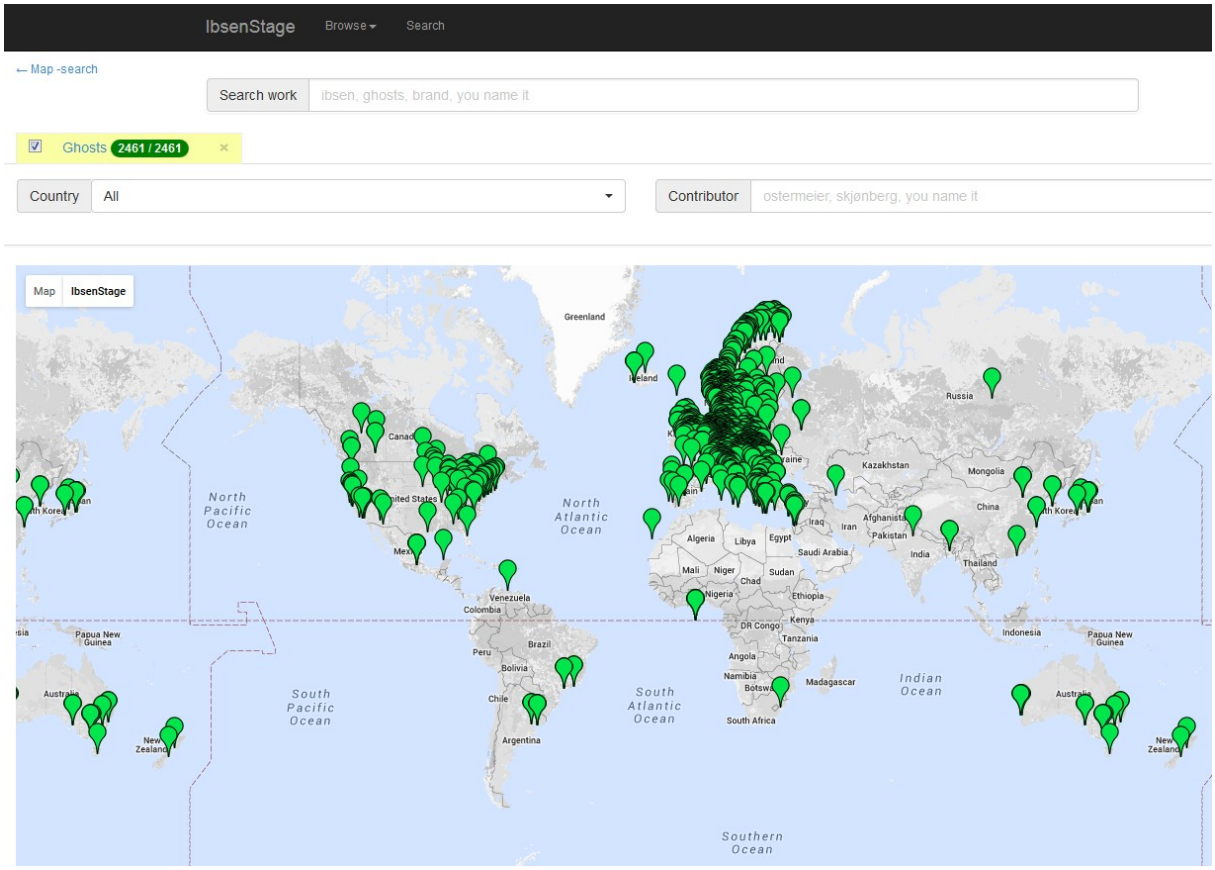
Henrik Ibsen - in text and on stage

Stine Brenna Taugbøl
Centre for Ibsen Studies, UiO
s.b.taugbol@ibsen.uio.no

The Centre for Ibsen Studies presents its digital resources on Ibsen: the performance database *IbsenStage* (ibsenstage.hf.uio.no) and the historical-critical edition of *Henrik Ibsen's writings* (ibsen.uio.no).

IbsenStage

IbsenStage is a performance database of Ibsen’s dramas reaching geographically all over the world and historically from when the dramas were first played and until today. The performances include theatre performances, radio, television and film productions. The web site has a map, a timeline and an advanced search. With these tools the users can delimit the results of their search to for instance work, country, director, actor, year or timespan. The database is continually updated with new performances. All information is in English.



The screenshot shows performances of *Ghosts* from 1882 until today.

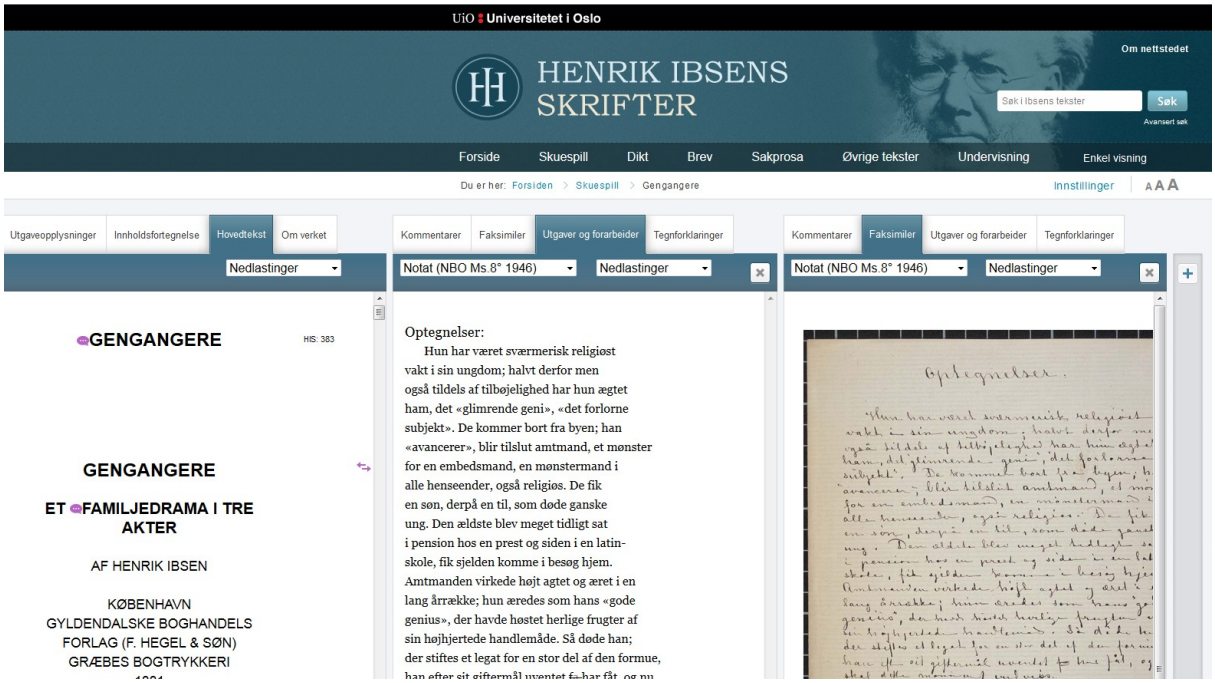
Contact: contact-ibsenstage@ibsen.uio.no
Technical support: DMLF, http://www.hf.uio.no/tjenester/it/hjelp/dmlf-dokumentasjon/

Henrik Ibsen's writings

Henrik Ibsen's writings includes all textual material after Ibsen (dramas, poems, prose, letters and varia) with introductions, commentaries and critical notes. The edition also offers a timeline of Ibsen's life and authorship, biographies on recipients and mentioned persons in letters, and exercises for schools and universities. The users can search in the text archive and delimit the results after genre, work, timespan or letter recipient.

Ibsen's texts are transcribed in their original language and spelling, while the editorial texts are in modern Norwegian.

The edition is available in two versions, a simple and an advanced version. The simple version gives only one commented and corrected version of every text, in most cases based on the first print. In the advanced version all versions of Ibsen's plays are available, that is both drafts and all printed editions in the author's lifetime. The different versions can be displayed side by side on the screen. In addition to this, facsimiles of all first prints and all manuscripts are available in the advanced version.



The main edition of Ibsen's drama *Ghosts* side by side with an early draft (in transcription and facsimile).

Contact: henrik-ibsens-skrifter@ibsen.uio.no
Technical support: EDD, http://www.hf.uio.no/iln/om/organisasjon/edd/

Utvikling av konseptet “LÆR KIDSA MUSIKKKODING”

Anders Lunde Tveit, Alexander Refsum Jensenius
Institutt for Musikkvitenskap, Universitetet i Oslo
a.r.jensenius@imv.uio.no

“Lær kidsa musikkkoding” er et prosjekt som springer ut av forskningsaktivitetene i fourMs-gruppen -- Music, Mind, Motion, Machines -- ved Institutt for Musikkvitenskap, Universitetet i Oslo, som jobber i grenselandet mellom opplevelse og skaping av musikk ved hjelp av ulike former for teknologier.

Målet med prosjektet er utviklingen av et pedagogisk konsept for design av elektronisk musikk med utgangspunkt i enkel programvare og elektronikk. I tillegg er vi opptatt av å vise at det er mulig å benytte musikk som utgangspunkt for å lære både matematikk, elektronikk og programmering for elever helt ned i grunnskolen. Prosjektet er dermed godt posisjonert i forhold til å svare på behovet for faglig fornyelse og konvergens som det etterspørres i rapporten *Fremtidens skole* fra Ludvigsen-utvalget [1].

Det har tidligere vært utviklet musikkprogramvare til bruk i videregående og grunnskole sammenheng i Norge men hvor fokuset har vært mer rettet inn mot komposisjon og lydbehandling ved bruk av lineærtid-basert sequencer programvare, slik som foreksempel *DSP* fra NoTam [2] og det EU/EØS initierte prosjektet *Composing With Sound*, CWS [3]. “Lær Kidsa musikkkoding” prosjektet har således en annerledes innfallsvinkel, hvor fokuset er rettet mot musikk/lydprogrammering og interaksjon, alt gjort i en nettleser.

Vi har gjennom flere år bygget opp både musikkteknologiemner som *Interaktiv musikk og Lydlære* og musikalsk samspillsopplegg som *Oslo Laptop Orchestra* og *Oslo Mobile Orchestra* som har vært rettet mot universitetsstudenter. Flere av ideene herfra er tenkt tilpasset og videreutviklet til bruk i skolesammenheng. Dette krever en annen skalering av både opplegg og teknologi, slik at det kan fungere med større grupper (20-30 elever) og innenfor andre tidsrammer (fra enkelttimer til ukeslange workshops).

Noen av de tekniske utfordringer en møter i musikkteknologi- og skolesammenheng er ofte ulike datamaskinparker, operativsystemer, samt sentralstyrt programvare-installering og drifting. Dette gjør at de mer tradisjonelle musikkprogrammeringsspråkene (Max, PureData, SuperCollider, CSound) ofte er vanskelige å tilpasse til bruk i en skolesammenheng.

Utviklingen av HTML5 i nye nettlesere og lydsyntese i sanntid gjennom standarden Web Audio API, har gjort det mulig å realisere musikalsk programmering direkte i en nettleser. Dette krever minimalt med drifts- og installeringsbehov, og fungerer på både datamaskiner og ulike mobilplattformer. I prosjektet «Lær kidsa musikkoding» valgte vi derfor å ta i bruk musikkprogrammeringsspråket *Gibber* [4] som bygger på nettopp disse nyeste nettleserteknologiene. Det har en enkel og forståelig syntaks (basert på Javascript), samtidig som det har et bredt bibliotek av ulike former for syntese, signalbehandling, sekvensering og sanntidslyd.

Den første fasen av prosjektet ble gjennomført høsten 2015 hvor deler av Gibber ble oversatt til norsk og tilpasset til det pedagogiske konseptet vi ser for oss. En annen viktig del av arbeidet har

vært å utvikle et sett musikalsk engasjerende eksempler, som også legger opp til en trinnvis økning i kompleksitet av matematikk, signalbehandling, og programmering. Dette har vi god erfaring med fra undervisning på universitetsnivå, og regner med at et tilsvarende konsept også vil kunne fungere i den videregående skolen.

Den neste fasen blir å teste ut opplegget på førsteårsstudenter ved Institutt for musikkvitenskap, og basert på en evaulering av dette vil vi deretter kjøre piloter på videregående skoler og kanskje også i grunnskolen.

[1] https://www.regjeringen.no/no/dokumenter/nou-2015-8/id2417001/
[2] http://www.notam02.no/web/1997/07/dsp/
[3] http://cws.dmu.ac.uk/EN/11
[4] http://gibber.mat.ucsb.edu

SEMATIA: Linguistic Annotation of Greek Papyri

Marja Vierros
University of Helsinki
marja.vierros@helsinki.fi

Greek papyri from Egypt have preserved bigger and smaller entities of Greek as it was written by the ancient speakers from ca. 300 BCE to 700 CE. There are different registers, styles and vernaculars. Therefore the papyrological corpus forms an important direct source for Greek linguists, which has been exploited very little due to the lack of good tools for searching linguistic structures. The papyrological corpus is freely available in digital form in the Papyrological Navigator platform (http://papyri.info/) which also allows users to search both text strings and metadata (such as date and provenance). The search possibilities do not, however, easily yield to searching linguistic structures or variation in spelling or morphosyntax.

It has been my goal for several years to develop a solid system for getting at least parts of the papyrological corpus linguistically annotated so that they would better serve scholars studying variation and change in Greek in the postclassical period. There were two bigger questions to be solved.

First, what kind of linguistic annotation would be most useful? Because there already are two linguistically annotated treebank corpora for Ancient Greek, the Ancient Greek Dependency Treebank (374,490 words of Homer, Hesiod, tragedies) and the PROIEL treebank (181,364 words of the New Testament and Herodotus),[1] the most reasonable solution was to follow the same framework (they both follow the Dependency Grammar) in order to gain maximal synergy between the corpora. Since a new annotation environment, Arethusa, was developed for the AGDT within the Perseids project (http://sites.tufts.edu/perseids/), my preliminary annotations of the papyri have followed the AGDT system.

The second question was more fundamental: How to deal with the fragmentary nature of the papyrological material within the annotation framework? The papyri are full of gaps and uncertain readings, abbreviations and nonstandard language which is not recognised by the automatic tools. For this purpose we have developed a layering tool called SEMATIA (http://sematia.hum.helsinki.fi) which filters two different layers from the EpiDoc XML format in which the texts are in the Papyrological Navigator. These layers are then ready for the linguistic

1See Haug 2014; Bamman & Crane 2011; Bamman et al. 2009; Haug et al. 2009.

annotation, one presenting only what was originally written and preserved in the papyrus and the other presenting the same text as a standardized version. By annotating both versions, it is possible to find the differences between standard and nonstandard language and search phenomena from both layers or only from the original layer. In this poster, I will present the layering tool and the current stage of annotations as well as discuss the querying possibilities and future developments we have been planning. Especially a third layer for marking up linguistic variation, and means for detecting and separating idiolectal features from scribal language are on the agenda.

References:

Bamman, David and Gregory Crane. 2011. "The Ancient Greek and Latin Dependency Treebanks." In Caroline Sporleder, Antal van den Bosch and Kalliopi Zervanou (eds.), *Language Technology for Cultural Heritage*, ser. Foundations of Human Language Processing and Technology (Springer), 79–98.

Bamman, David, Francesco Mambrini, and Gregory Crane. 2009. "An Ownership Model of Annotation: The Ancient Greek Dependency Treebank" .

Haug, Dag 2014. "Computational Linguistics and Greek." Encyclopedia of Ancient Greek Language and Linguistics. Edited by: Georgios K. Giannakis. Brill Online, 2014.

Haug, D.T.T., Eckhoff, H. M., Majer, M., Welo, E. 2009. "Breaking down and putting back together: analysis and synthesis of New Testament Greek." *Journal of Greek Linguistics* 9, 56-92.

Repertory of Conjectures to Horace

Tor Ivar Østmoe
University of Oslo
t.i.ostmoe@ifikk.uio.no

In the textual criticism of Greek and Roman authors, a conjecture is an educated guess indended to improve sense in a passage that in the manuscripts is incomprehensible or unsatisfactory. Horace (65-8 BC) is a canonical Roman poet whose poems are extant in c. 250 manuscripts. A larger number of conjectures have been proposed on Horace than on other authors. *Repertory of Conjectures to Horace* (http://tekstlab.uio.no/horace/intro) is a database at the University of Oslo that lists and analyses conjectures according to type and provides bibliographical information. This information also includes subsequent scholarship. The poster / presentation presents the database and its chosen format.

Pre and post conference workshops

Digital Classics. A hands-on Introduction to EpiDoc and Treebanking.

Organizer: Federico Aurora
University Library of Oslo/Unversity of Oslo
federico.aurora@ifikk.uio.no

This workshop will be a practical introduction to a set of resources, methodologies and practices, centered around the digital annotation of texts, which in the last decades have come to enrich, and partially change, the field of classical studies (Crane & Terras 2009), especially – but not only – in the areas of papyrology (Depauw & Stolk 2015), epigraphy (Cayless et al. 2009), linguistics (Bamman 2014; Haug 2014) and textual criticism (Dué & Ebbott 2009). There will be two sessions, one on text encoding with EpiDoc and one on treebanking for ancient languages.

The workshop being an introduction, it will firstly be directed to scholars and students with little or no experience in using TEI/EpiDoc or with treebanking, but, of course, also to more advanced participants interested in the use of TEI encoding and treebanking with ancient languages.

Maximum number of participants: 20

Time and Location: 9.00-13.00, Monday 14.03.2016, "Læringsoasen" room, Humanities and Social Sciences Library, University Campus Blindern

Technical equipment: None. The room will be equipped with a desktop computer (with the necessary software already installed) for each participant.

Organizers: Humanities and Social Sciences Library (UiO), Department of Philosophy, Classics, History of Art and Ideas (UiO).

Contact: Federico Aurora, federico.aurora@ub.uio.no

Session 1: EpiDoC

EpiDoc[1] is a set of guidelines for encoding the digital versions of ancient source texts in TEI (Text Encoding Initiative)[2] – originally developed for Greek and Roman epigraphy, but now much more diverse – including a recommended schema and ODD (One Document Does it all) file,[3] a lively community of practice, and an ecosystem of projects, tools and stylesheets for the interchange and exploitation of such texts (Bodard 2010).

The first workshop session will introduce participants to the principles and practices of EpiDoc encoding, which are largely based on the practice of encoding single-source documents and the ancient objects on which they are written. Special attention will be paid to how the community guidelines recommend (a) mapping Leiden Conventions for text transcription (i.e. the standard practice for paper editions, see Cayless et al. 2009: 9) to TEI transcription and critical apparatus elements, and (b) mapping the descriptive and historical metadata features – from projects such as

1 epidoc.sourceforge.net
2 http://www.tei-c.org/index.xml
3 http://wiki.tei-c.org/index.php/ODD

EAGLE (Europeana network of Ancient Greek and Latin Epigraphy)[1] and Papyri.info[2] – to TEI "MSDesc" (manuscript description) elements.

Finally, tools and other methods made available by the community for transforming, publishing, querying, exchanging, and linking encoded materials will also be introduced.

Instructors:

Gabriel Bodard, reader in digital classics at the Institute of Classical Studies, University of London, has a background in classics, with training and experience in both papyrology and epigraphy. He worked for fourteen years in Digital Humanities at King's College London on various digital projects and acquired broad and in-depth knowledge of text encoding and processing, especially the use of TEI XML and XSLT. He is one of the lead authors of the EpiDoc Guidelines (Elliot, Bodard et al. 2007-2014) and regularly organises and teaches training workshops in digital epigraphy and papyrology.

Joanne Stolk, PhD fellow at the Department of Philosophy, Classics, History of Art and Ideas, University of Oslo, has written her PhD thesis on "The Language of the Papyri: Diachronic change in Egyptian Greek (300 B.C. - 800 A.D.)". She has broad experience with encoding papyrus texts in the Papyrological Editor of Papyri.info and has contributed to create a database on irregularities and mistakes in the Greek papyri (Trismegistos Text Irregularities).[3]

Session 2: Treebanking

"A treebank is a large collection of sentences in which the syntactic relation for every word is made explicit – where a human has encoded an interpretation of the sentence in the form of a linguistic annotation." (Bamman 2015). Treebanking means, thus, annotating texts morpho-syntactically (and, more recently, also semantically and for information and discourse stucture). This allows to approach the study of ancient languages, and especially of their more abstract structures, in more sophisticated ways, developing empirical analyses and testing theoretical assumptions against large amounts of data in a statistically relevant way. Diachronic treebanks, moreover, especially of languages for which we can count on corpora stretching over more than two thousands years, such as Greek and Latin (including also its Romance continuators), also offer an invaluable opportunity to perform quantitative studies of language change and evolution. And even though in its beginnings the creation of treebanks regarded mostly modern languages, the last years, which have seen a large growth of corpora for ancient languages, (some of them quite general and available online, others created ad hoc and for specific research projects) have also seen the creation of treebanks of historical languages, such as the The Ancient Greek and Latin Dependency Treebank (Bamman & Crane 2007; Bamman et al. 2009) at the Perseus Project[4] and the PROIEL Treebank (Haug & Jøhndal 2008), currently the two largest treebanks of Ancient Greek and Latin.

The second workshop session will give a basic introduction to the theory, methodology and practice of treebanking seen in the light of the particular challenges posed to it by ancient languages (focusing especially on Greek and Latin). Its practical part will be a hands-on demostration of the PROIEL treebank and its online application.[5] The PROIEL treebank, originally developed for the PROIEL project (Pragmatic Resources in Ancient Indo-European Languages) to contain the original Greek text of the New Testament and, aligned, its translations into early Indo-European languages (Gothic, Armenian, Latin and Old Church Slavic), has later been extended with several other texts in Greek and Latin (e.g. Herodotus, Caesar, Cicero, the *Peregrinatio Aegeriae*),

1 http://www.eagle-network.eu
2 http://papyri.info
3 http://www.trismegistos.org/textirregularities/
4 http://www.perseus.tufts.edu/hopper/
5 http://foni.uio.no:3000

but also in other historical languages, including the Old Norse *Edda* and many texts in old Romance languages.

Instructor:

Dag Haug, professor of Latin at the Department of Philosophy, Classics, History of Art and Ideas, University of Oslo, works in historical linguistics, syntax (Lexical Functional Grammar) and formal semantics, having his background in Homeric philology and comparative Indo-European linguistics. He led the PROIEL project and the construction of the PROIEL treebank.

Person responsible for the workshop:

Federico Aurora, subject specialist in classics and papyrus collection curator at University of Oslo Library and PhD candidate at the Department of Philosophy, Classics, History of Art and Ideas, University of Oslo, is writing a doctoral thesis on the case system of Mycenaean Greek and leads the construction of the annotated corpus of Mycenaean Greek (DAMOS - Database of Mycenaean at Oslo).[1]

The workshop will be followed in the afternoon by a lecture by Gabriel Bodard with the title:

"A Digital Classics research agenda: both Digital Humanities and Classical Studies".

The lecture will be hosted by the Classical Seminar of the Department of Philosophy, Classics, History of Art and Ideas, and will be open to everybody.

Time and place for the lecture: *Monday March 14, 2016, 14:15-16:00, room 207, Georg Morgenstiernes building, Blindern University Campus.*

References

Bamman D., 2014. D. Bamman. "Dependency Grammar and Greek." Encyclopedia of Ancient Greek Language and Linguistics. Brill Online, 29 November 2015. First appeared online: 2014. <http://referenceworks.brillonline.com/entries/encyclopedia-of-ancient-greek-language-and-linguistics/dependency-grammar-and-greek-COM_00000091>

Bamman & Crane 2007. “The Latin Dependency Treebank in a cultural heritage digital library”. In: C. Sporleder, A. van den Bosch, C. Grover, eds. *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)* , 33–40. Prague.

Bamman, D., F. Mambrini & G. Crane. 2009. “An ownership model of annotation: the Ancient Greek Dependency Treebank”. In: M. Passarotti et al., eds. *The Eighth International Workshop on Treebanks and Linguistic Theories (TLT 8)*, 5-17. Milan.

Bodard, G. 2010. "EpiDoc: Epigraphic documents in XML for publication and interchange." In: F. Feraudi-Gruénais, ed. *Latin on Stone: Epigraphic Research and Electronic Archives,* Roman Studies: Interdisciplinary Approaches. Rowan & Littlefield, 101-118. Available:

1 https://www2.hf.uio.no/damos/

http://www.stoa.org/wordpress/wp-content/uploads/2010/09/Chapter05_EpiDoc_Bodard.pdf

Cayless, H., C.M. Roueché, E. Elliot, G. Bodard 2009. "Epigraphy in 2017." In: Crane & Terras 2009
Crane. G., M. Terras, eds. 2009. *Changing the Center of Gravity: Transforming Classical Studies Through Cyberinfrastructure. Digital Humanities Quarterly*, 3, 1. Available: http://www.digitalhumanities.org/dhq/vol/3/1/index.html

Depauw, M. & J. Stolk 2015. "Linguistic Variation in Greek Papyri: Towards a New Tool for Quantitative Study." *Greek, Roman and Byzantine Studies*, 55, 1, 196-220. Available: http://grbs.library.duke.edu/article/view/15245/6561

Dué, C., M. Ebbott 2009." Digital Criticism: Editorial Standards for the Homer Multitext". In: Crane & Terras 2009.

Elliot, Bodard et al. 2007-2014. Elliott, T., G. Bodard, E. Mylonas, S. Stoyanova, C. Tupman, S. Vanderbilt, et al. *EpiDoc Guidelines: Ancient documents in TEI XML (Version 8)*. Available: http://www.stoa.org/epidoc/gl/latest/.

Haug, D 2014. "Computational Linguistics and Greek." *Encyclopedia of Ancient Greek Language and Linguistics*. Brill Online, 2015. 30 November 2015. First appeared online: 2013. <http://referenceworks.brillonline.com/entries/encyclopedia-of-ancient-greek-language-and-linguistics/computational-linguistics-and-greek-COM_00000072>

Haug, D. T. T. & Jøhndal, M. 2008. "Creating a parallel treebank of the Old Indo-European bible translations". In: K. Choukri, ed., *Proceedings of the second workshop on language technology for cultural heritage data (LaTeCH 2008)*. Available: http://www.hf.uio.no/ifikk/english/research/projects/proiel/Activities/proiel/publications/marrakech.pdf

Bukvik - a DH Scholar's Environment for Stylistic Analysis

DHN Workshop

Organizers:
Sasha Mile Rudan (Oslo University, sasharu@ifi.uio.no)
Eugenia Kelbert (University of Passau, eugenia.kelbert@uni-passau.de)
Approximate Number of Participants: 20+

Approximate Schedule
9 am-9:50 am - Presentation of recent advances in stylometry and overview of available tools and their capacities, pros and cons etc. Presentation of Bukvik and its functions. Setting up experiments in Bukvik on individual computers, with the help of the organizers. Participants are encouraged to bring their own research corpora to analyse if they wish and

will be provided instructions on the formatting prior to the workshop. **9:50 am-10 am** - Break and informal Q&A
10 am-10:50 am - Running experiments. Obtaining, visualizing and sharing results. **10:50 am-11 am** - Break and informal Q&A
11 am-noon - Running experiments.Obtaining, visualizing and sharing results. General discussion.

Proposal

Digital Humanities, in literary studies, are often associated with ***distant reading***, Franco Moretti's term for tackling enormous corpora of fiction and poetry. With this method, it becomes possible to **trace tendencies**, notably, in the development of genre that would otherwise escape a researcher's notice. Many scholars familiar with distant reading view it as an alternative to **close reading**, a more traditional critical approach. For some, this leads to **skepticism** - surely, no machine is powerful enough to substitute for the training and the nuanced eye of a literary scholar?

A number of recent publications, however, mark a shift of awareness. For example, two monographs, Matthew Jockers' *Macroanalysis* and David Hoover's and his colleagues' *Digital Literary Studies*, represent a turn towards a focus on the potential of the Digital Humanities when it comes to stylistic analysis. Both studies exemplify a growing awareness that computational methods known as **stylometry** (Jockers, Rybicki) or computational stylistics (Hoover) may already be capable of what Jockers refers to as a quantitative approach that "allows for both **zooming in and zooming out**" (22).

In other words, computational tools today have reached a stage where they may indeed contribute to our understanding of **style**. While computers cannot read in the same way as a human does, they can provide that human with highly relevant information. **Tools** such as Part-of-Speech tagging, semantic analysis, or some of the methods developed originally to aid authorship attribution provide data that **no close reader can obtain** but that may, especially when gathered with an eye to stylistics rather than categorization, be fruitfully used by a close reader. Hence, we may now speak of a potential, albeit limited, ***stylistic profile of a corpus***: the sum total of quantifiable features for each text or body of texts that, together, constitute a multidimensional fingerprint of the given writer's style with reference to a balanced corpus of fiction in the given language.

This workshop will include a brief overview of **methods and tools** currently available in stylometry, and of their application in literary analysis. It will, further, introduce participants to a new tool called **Bukvik**, designed specifically to aid stylistic analysis across languages. Bukvik is an online environment for DH scholars to map their research experiments/processes through a visual data-flow diagramming interface that gives them access to different tools ranging from purely linguistic and semantic to tools supporting comparative and network analysis. Apart from classical NLP support such as Part-of-Speech tagging, it provides **knowledge extraction** and **(S)NA** support (that help us describe and understand relations between document artifacts; keyword markers, characters, locations, topics), and building and analyzing stylistic profiles (the totality of quantifiable features in a text that can be traced computationally and conceived as a multidimensional space that can be compared with a similar space with the same parameters of other author, piece, or translation).

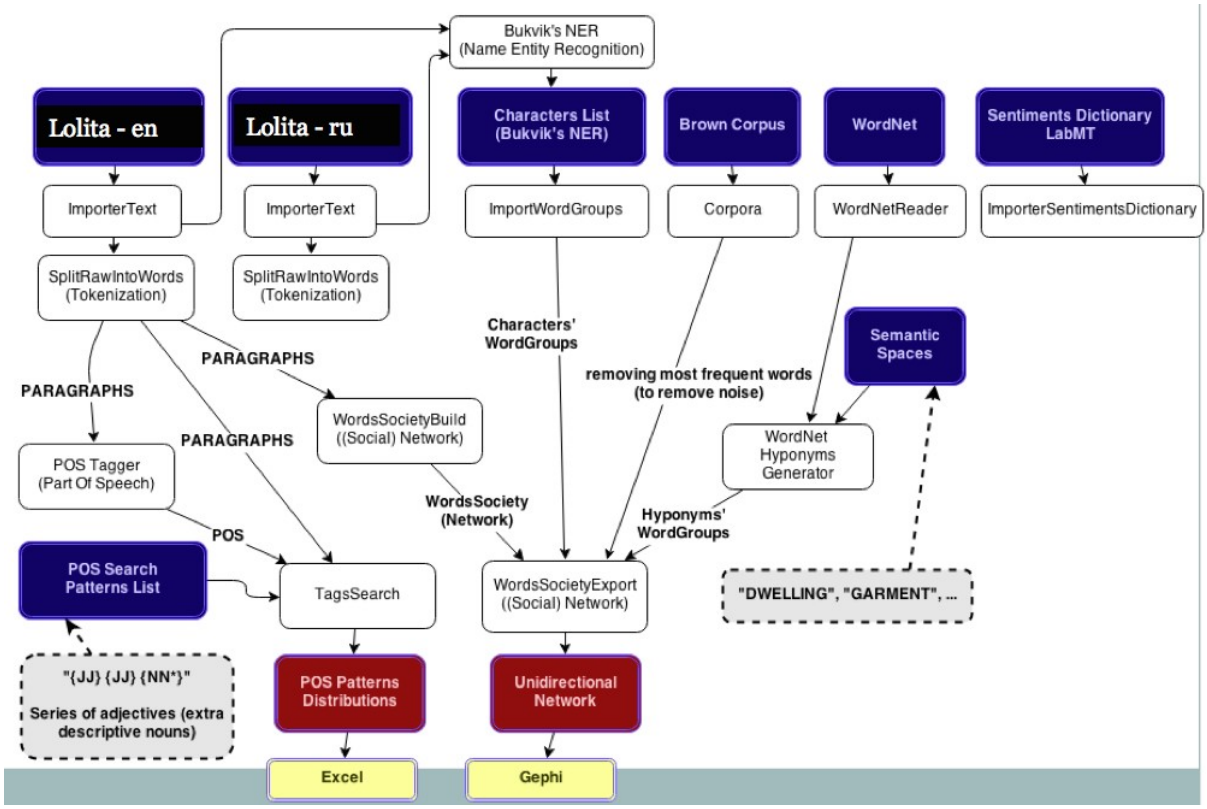


Figure 1: Conceptual diagram of a research flow in Bukvik

The system is a comprehensive environment for scholars in literary studies, sociology, history etc., interested in considering and comparing stylistic features of large corpora, potentially across languages. It has a modular structure and is able to include existing state-of-the-art external tools. Thus, **without any prior knowledge of programming** or having to switch between a variety of different systems, formats and interfaces, the user is able to customize the experiment to suit his or her own research task. On the other hand, more advanced user may create new modules and tools written in Python, JavaScript, and other programming languages and frameworks. As a research experiment is built, the user is able to **select which tools to apply** and to see the experiment visualized automatically as a user-friendly flow (***CollaboFlow***), enabling experiment **sharing**, commenting and online collaboration. Finally, results may be **visualized** as diagram; the tool also generates **examples** and allows the user to access them within the original text, so as to inform distant reading with a close reading approach and vice versa. Bukvik can be applied, among other fields, to the study of bilingual authors, translation or the comparison of articles of large corpora such as Wikipedia written in different languages. The workshop is taught by the **developers of Bukvik** and it is hoped that it may lead to **future collaborations** as further applications of the system, applied to a variety to texts and research projects, become evident during the session. The **participants** are therefore welcome to get in touch with the organizers and provide their **own corpora** prior to the conference, which they may use to practice using the system as applied to their own research already during the workshop. This, however, is optional, as we will also offer **our own corpora** to play with.

The system is conceived as an **aid for automatic zooming**: unlike the "distant reading" approach where statistics replaces reading and helps process large corpora, we see Bukvik as an automatic framework that will ultimately aid and direct close reading. "Close reading at a distance" or **Qualitative Augmented Quantitative Analysis** is one way to describe the idea behind the methodology. The goal is for the two approaches to interact and inform each other: qualitative data will shape and instruct the quantitative component in analysis leading to more relevant results. Having this flexibility, Bukvik allows scholars a **variety of experiments**, such as the analysis and

differential parallel comparison of translations of the same book, of an original with a translation, of corpora of two writers' work, as well as comparing texts within a language or across languages, and comparing variations from respective corpora in each language.

Organizers

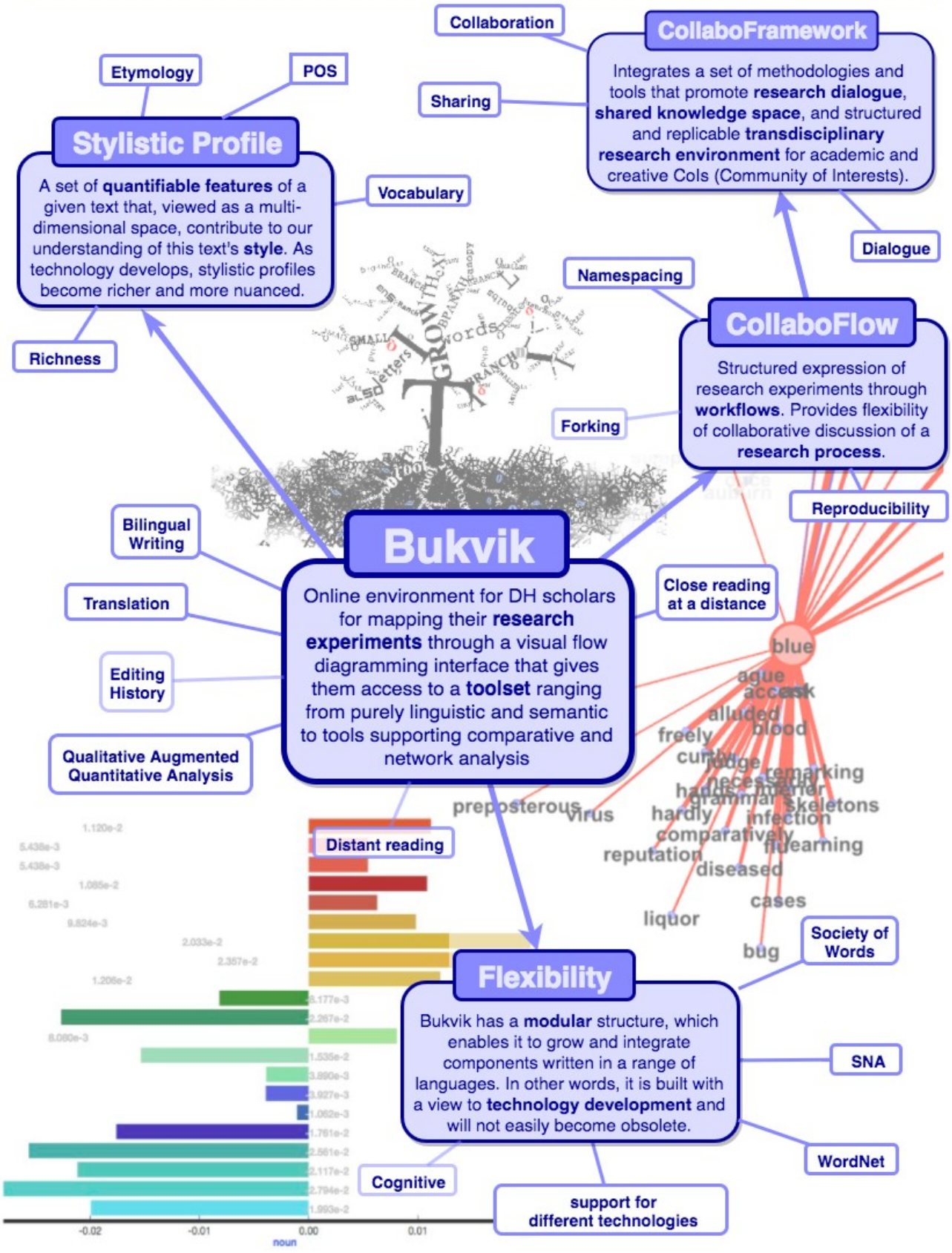
Institutions supporting the workshop: Oslo University, ChaOS (Cultural Humane Art Of Science), and Centar za digitalne humanističke nauke (CDHN) Belgrade.

Saša Mile Rudan is completing his Ph.D. in Computer Science at Oslo University. He is affiliated with Serbian DH center and cofounder of ChaOS organization. He specializes in complex collaborative socio-technical systems (be it architecting and conducting research on them, leading transdisciplinary production teams, or taking an entrepreneurial role), social processes, and knowledge management. His interest in literary analysis lies in bridging the socio-technical gap for scholarly research, in the Qualitatively Augmented Quantitative Analysis (QaQa) methodology, cross-lingual and cross-cultural comparison and in supporting under-resourced languages. These interests led to his involvement in projects: LitTerra (www.LitTerra.info; augmenting books, providing a literary-ecosystem with deep in-book and inter-book insights) and Bukvik (Bukvik.LitTerra.info), a research infrastructure for literary scholars.

Eugenia Kelbert is a Lecturer at Passau University in Germany. She completed her PhD in Comparative Literature at Yale University in May 2015. Her current book project is based on her dissertation, and focuses on the phenomenon of literary translingualism in the 20th and 21st century, and included extensive quantitative results on bilingual writers' work in different languages. She has written on Brodsky, Nabokov and Romain Gary, as well as other translingual writers. Her interests include Digital Humanities, translation theory, stylistics, national and identity discourses, poetry, and quantitative approaches to textual analysis.

Bukvik

a DH Scholar's Environment for Stylistic Analysis



CollaboFramework

Framework and Methodologies for Collaborative Research in Digital Humanities

DHN Workshop

Organizers:
Sasha Mile Rudan (Oslo University, sasharu@ifi.uio.no)
Sinisa Rudan (Belgrade University, sinisa.rudan@gmail.com)
Dino Karabeg (Oslo University, dino@ifi.uio.no)
Eugenia Kelbert (University of Passau, eugenia.kelbert@uni-passau.de)
Approximate Number of Participants: 20+

Approximate Schedule
30 min **D+H = DH**: The needs of transdisciplinary collaborative research. Challenges and benefits from merging science and the humanities, merging their methodologies and knowledge bases.
30 min **CollaboFramework**: Presenting concepts and methodologies for transdisciplinary collaborative research. Presenting DH related tools for collaboration and structured research. 15 min Break and informal Q&A
45 + 30 min **Collabo-in-vivo (45)**: Practicing transdisciplinary collaboration with **CollaboFramework** (a set of methodologies and tools). Practicing at-place (located) vs. dislocated collaboration and real-time vs. asynchronous (offline) collaboration. **research-flow (30)**: Practicing analyzing research experiments, visualizing and sharing results with ***CollaboFlow*** (on the example of ***Bukvik***).
Last 30 min **dialogue * dialogue = dialogue**: discussion on the fundamentals of transdisciplinary collaboration, its sustainability, improvements and systemic approach.

Workshop Info

This workshop addresses the needs and challenges of collaboration in transdisciplinary research, as presented through the use case of Digital Humanities.
The very nature of Digital Humanities calls for interdisciplinary collaboration between the Humanities and science. The collaborative process requires assimilating across the digital-humanities gap, but also a socio-technical gap. These two academic communities are distant on the academic landscape, which very often results in misunderstanding, causes difficulties in communication and can even reduce the quality or correctness of research results. During the workshop, we present **problems, possible approaches and research scenarios** in DH and interdisciplinary research in general. We present concrete usecases

and guide participants through the collaborative research process.

CollaboFramework (CF) integrates a set of methodologies and tools that (among other aspects) promote **research dialogue**, a **shared knowledge space**, and a **structured and replicable transdisciplinary research environment**.

We utilize CollaboFramework for addressing the aforementioned requirements and challenges. CollaboFramework is aiming to create a collaborative space for transdisciplinary academic and creative **CoI** (Community of Interests) in which it will - *through continuous research dialogue, co-creation, and mediating mechanisms* - garden the interdisciplinary research process. Through this process, CollaboFramework helps to evolve interdisciplinary understanding, knowledge, creative solutions, and concrete actions. ***CollaboFramework*** is methodologically forked for different domains - ***CollaboScience*** and ***CollaboArte***, among others.

Technology-mediated continuous change of collaboration space - *CollaboDialogue*

Understanding the importance of **dialogue as defined in David Bohm's work** and as a means of an incremental mechanism of collaborative research, we recognize dialogue as a **paradigm shifter** in DH research. In inheriting the semiotic essence of the term dialogue, it calls for a continuous change of collaboration space, support for evolutional rather than revolutionary changes, or in other words, a mechanism for non-intrusive, but rather constructive changes. We practice this concept across the whole framework, affecting, among others, knowledge production, shared understanding, research structuring and research evaluation. CollaboDialogue being a paradigm shifter, it helps us in grounding principles for **transformation of the very socio-technical cyberinfrastructure used for DH tools**

"CollaboFramework aims for low-friction and non-disruptive mechanisms for performing the sustainable loop of research-dialogue (Figure 1)".

In that way, practicing both analytical and synthetical phases of collaborative dialogue, we are capable of continuous creating/collecting and curating data and finally converging toward knowledge and insights, while we are iterating through the loop.
Through the workshop, we will present this iteration and its benefits by using **prepared examples**. We will demonstrate both ***at-place (located) vs. dislocated*** and ***real-time vs. asynchronous (offline)*** collaboration dialogue.

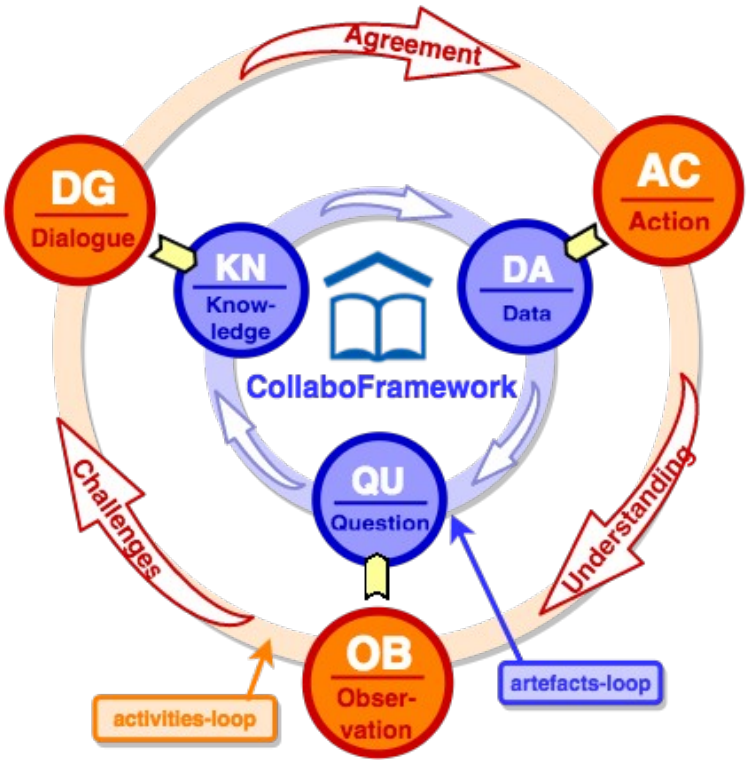


Figure 1: Principles of ***CollaboFramework - the spiral of sustainable research collaboration***

KnAllEdge - a Collective Knowledge space
KnAllEdge is a knowledge layer that resides under the CollaboFramework and unites all the components of the framework in a low-friction collective knowledge space. In this sense, CF tools are continuously extending, modifying and improving the collective knowledge space, rather than producing isolated heterogeneous artefacts that would require constant additional maintenance work (articulation work). In order to support transdisciplinarity, ambiguity and convergence of shared understanding, KnAllEdge support fuzziness, multi-truth/opinion, and semantic richness and extendability. Through a set of extensions, it supports semantic searching, argumentation (dialogue mapping), boundary objects, RIMA (community interests mapping and filtering), real-time chatting, real-time knowledge management moderation, voting, namespacing, notifications, etc.

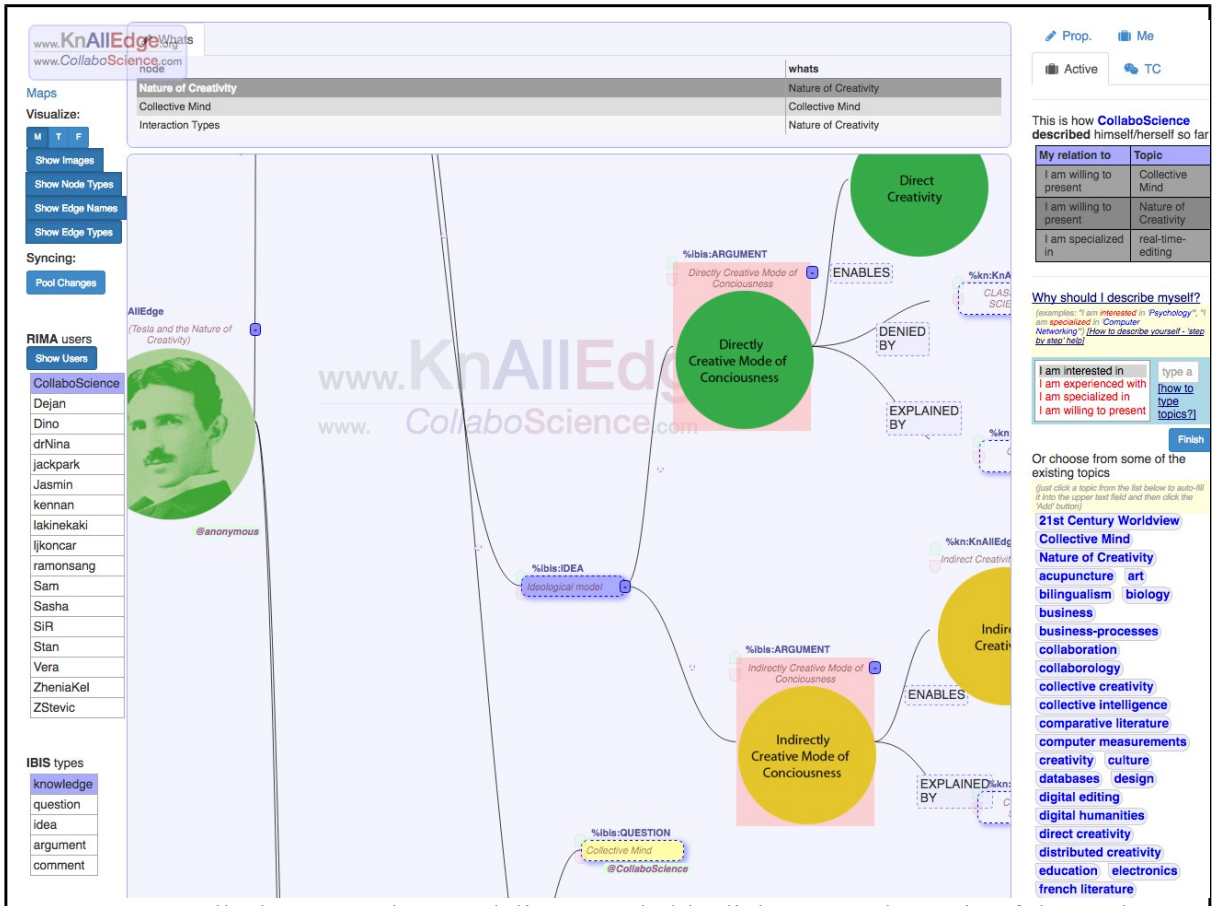


Figure 2: KnAllEdge example - Modeling roundtable dialogue on the topic of the Tesla's Nature of Creativity at the Tesla Conference (2015)

CollaboFlow - structured expression of research experiments

CollaboFlow - is a component that promotes structured expression of research experiments through **workflows** that are capable of capturing different aspects of a research experiment, while providing flexibility of collaborative discussion of a research process, proposing changes and eventually creating/forking research experiment replications and versioning of both experiment and datasets. In this way, it supports **research reproducibility** and encourages falsifiability.

Bukvik - an online environment for DH scholars

During the workshop, we will **demonstrate** it through real examples of stylistic research experiments of bilingual writers, realized with ***Bukvik*** - an online environment for DH scholars interested in textual (stylistic) corpora analysis (and the subject of a pre-conference workshop at DHN). Bukvik integrates CollaboFlow for experiment design together with presentation and collaborative discussion of research outcomes.

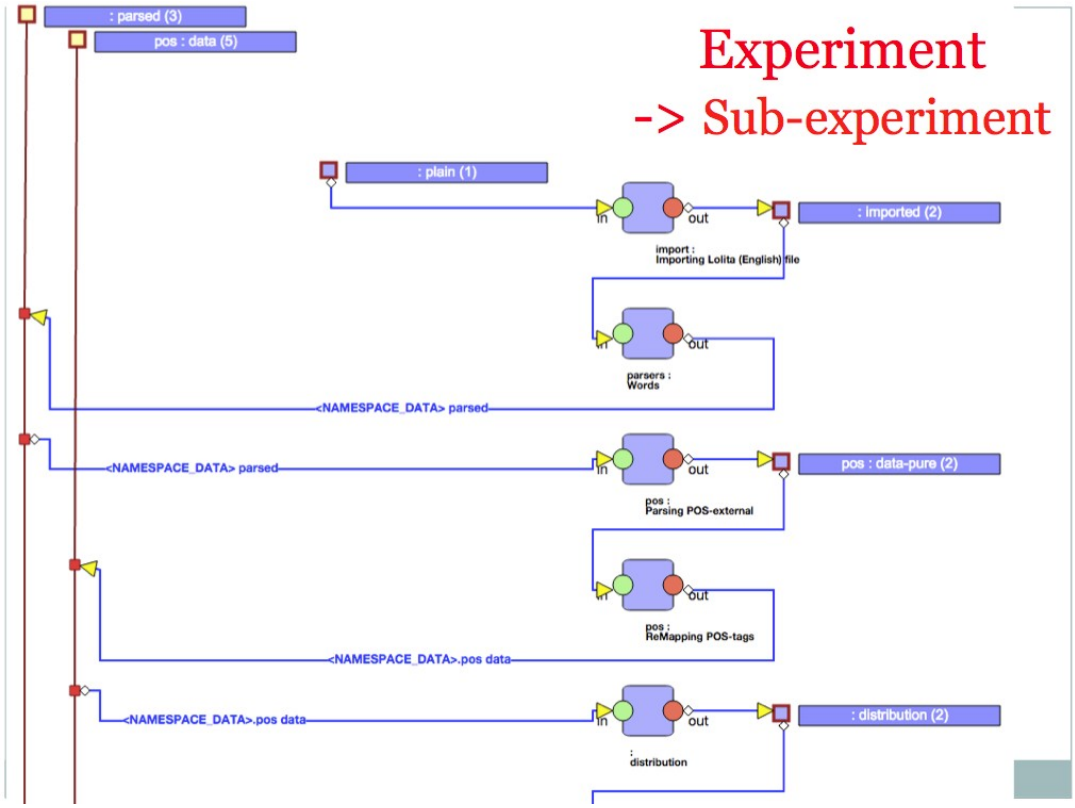


Figure 3: Bukvik example - Stylistic analysis of Nabokov's fictional corpora

Organizers

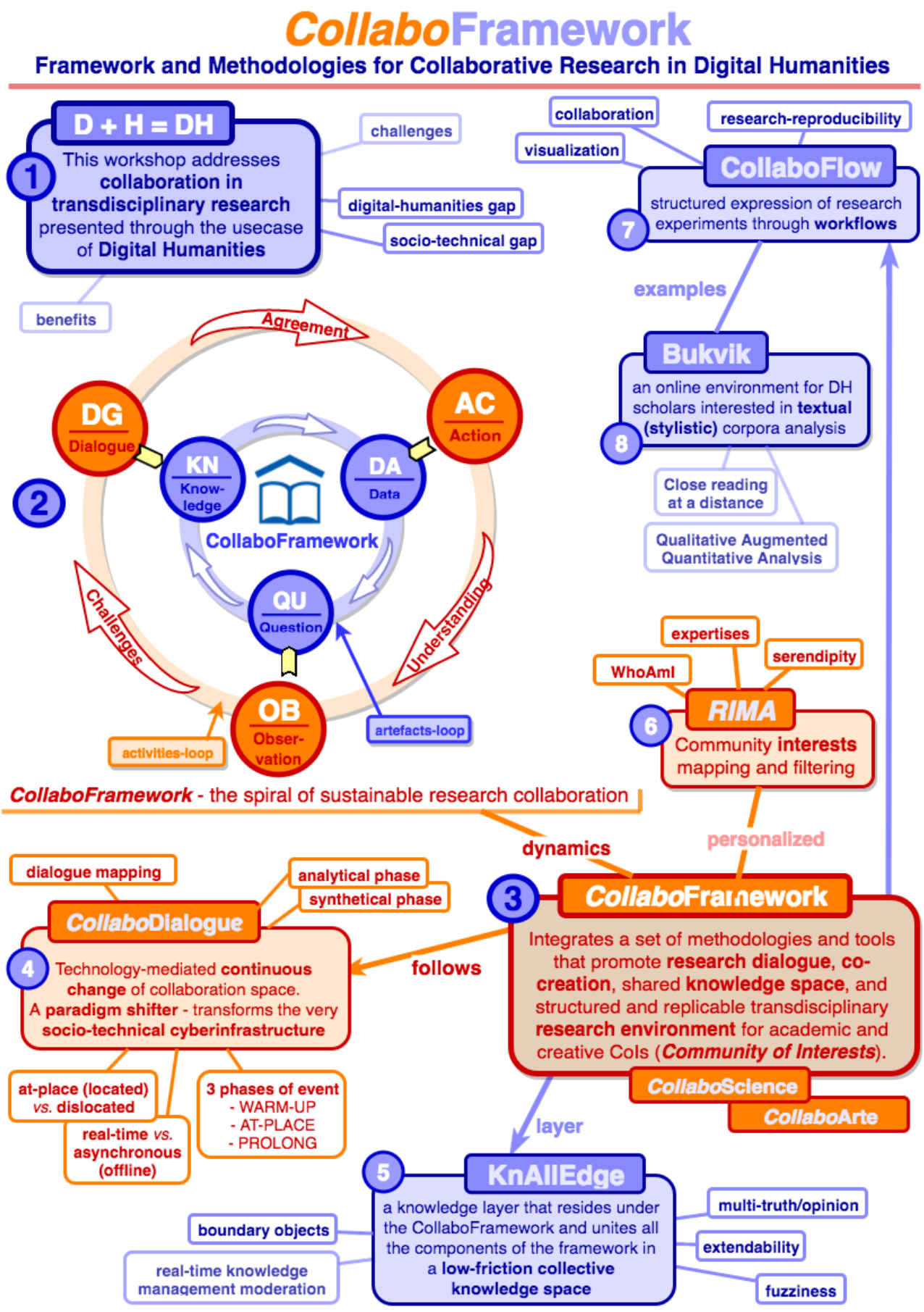
Institutions supporting the workshop: Oslo University, ChaOS (Cultural Humane Art Of Science), Knowledge Federation, and Centar za digitalne humanističke nauke (CDHN) Belgrade.

Saša Mile Rudan is completing his Ph.D. in Computer Science at Oslo University. He is affiliated with Serbian DH center and cofounder of ChaOS organization. He specializes in complex collaborative socio-technical systems (be it architecting and conducting research on them, leading transdisciplinary production teams, or taking an entrepreneurial role), social processes, and knowledge management. His interest in literary analysis lies in bridging the socio-technical gap for scholarly research, in the Qualitatively Augmented Quantitative Analysis (QaQa) methodology, cross-lingual and cross-cultural comparison and in supporting under-resourced languages. These interests led to his involvement in projects: LitTerra (www.LitTerra.info; augmenting books, providing a literary-ecosystem with deep in-book and inter-book insights) and Bukvik (Bukvik.LitTerra.info), a research infrastructure for literary scholars.

Sinisa Rudan is cofounder of ChaOS organization and owner of MagicWand Solutions. He is an independent researcher in the domain of collaboration, social-psychology in socio-technical systems, and leadership; an entrepreneur and an artist. He was coordinator for academic community of the The National Council for Serbian Language and Script and on several positions in lecturing, editing specialized magazines and conferences, jurying competitions in the area of multimedia and IT, as well as leading teams on transdisciplinary projects. His work and research is focused on creation of methodologies and tools for transdisciplinary collaboration of arts, science, and humanities professionals.

Dino Karabeg is a professor at the Oslo University Department of Informatics. With a PhD in computer science and a MS in environmental system analysis, he has been active in the area of complex systems and knowledge representation for more than 20 years. He is the creator of the concept and methodology of "*Knowledge Federation*" (http//www.knowledgefederation.org) a

sound framework designed to integrate disparate worldviews into coherent and operational systems.
Eugenia Kelbert is a Lecturer at Passau University in Germany. She completed her PhD in Comparative Literature at Yale University in May 2015. Her current book project is based on her dissertation, and focuses on the phenomenon of literary translingualism in the 20th and 21st century, and included extensive quantitative results on bilingual writers' work in different languages. She has written on Brodsky, Nabokov and Romain Gary, as well as other translingual writers. Her interests include Digital Humanities, translation theory, stylistics, national and identity discourses, poetry, and quantitative approaches to textual analysis.


CollaboFramework
Framework and Methodologies for Collaborative Research in Digital Humanities
D + H = DH
1
This workshop addresses collaboration in transdisciplinary research presented through the usecase of Digital Humanities
challenges
digital-humanities gap
socio-technical gap
benefits
collaboration
visualization
research-reproducibility
CollaboFlow
7
structured expression of research experiments through workflows
examples
Bukvik
8
an online environment for DH scholars interested in textual (stylistic) corpora analysis
Close reading at a distance
Qualitative Augmented Quantitative Analysis
Agreement
DG
Dialogue
AC
Action
KN
Know-ledge
DA
Data
2
CollaboFramework
QU
Question
OB
Obser-vation
Challenges
Understanding
activities-loop
artefacts-loop
expertises
serendipity
WhoAmI
RIMA
6
Community interests mapping and filtering
CollaboFramework - the spiral of sustainable research collaboration
dynamics
personalized
dialogue mapping
analytical phase
synthetical phase
CollaboDialogue
4
Technology-mediated continuous change of collaboration space.
A paradigm shifter - transforms the very socio-technical cyberinfrastructure
follows
3
CollaboFramework
Integrates a set of methodologies and tools that promote research dialogue, co-creation, shared knowledge space, and structured and replicable transdisciplinary research environment for academic and creative CoIs (Community of Interests).
CollaboScience
CollaboArte
at-place (located) vs. dislocated
real-time vs. asynchronous (offline)
3 phases of event
- WARM-UP
- AT-PLACE
- PROLONG
layer
5
KnAllEdge
a knowledge layer that resides under the CollaboFramework and unites all the components of the framework in a low-friction collective knowledge space
boundary objects
real-time knowledge management moderation
multi-truth/opinion
extendability
fuzziness

Author Index

Almevik, Gunnar 47
Ameel, Lieven 48
Aspaas, Per Pippin 60
Aurora, Federico 49, 159, 161
Battista, Simonetta 142
Baunvig, Katrine 52
Bergenmar, Jenny 51
Birkenes, Magnus Breder 27, 29
Biström, Anna 54
Blümm, Mirjam 135
Bodard, Gabriel 160
Boer, Johan von 90
Bognerud, Magnus 55
Borin, Lars 117, 122
Bree, Pim van 81
Brügger, Niels 56
Carbon, Jan-Mathieu 110
Clareborn, Anna 62
Coats, Steven 57
Cock, Coppélie 58
Conzett, Philipp 60
Culeddu, Sara 61
Cullhed, Per 62
Dagson, Trausti 87
Dagsson, Trausti 63
Domeij, Rickard 136
Eide, Oda Bakken 65
Eide, Øyvind 65
Eisenhauer, Monika 21, 24
Ekman, Stefan 122
Eriksson, Maria 67
Esborg, Line 19
Fridlund, Mats 67
Gjesdal, Anje Müller 68
Grufstedt, Ylva 70
Grønvik, Oddrun 72
Hagen, Kristin 141
Hanssen, Jens-Morten 73
Haug, Dag 161
Haugen, Trond 27, 29
Helland, Frode 75
Henriksson, Erik 137
Hietanen, Mika 73
Holledge, Julie 75
Holmen, Jon 72
Holownia, Olga 76
Homan, Zenobia 139
Jarlbrink, Johan 139

Jensenius, Alexander Refsum 77, 147, 154
Johannessen, Janne Bondi 141
Johannsson, Ellert 142
Johansson, Anna 58
Johnsen, Lars G. 27, 29
Jongejan, Bart 144
Jordan, Caspar 122
Kahila, Maarit 48
Kanner, Antti 78
Karabeg, Dino 167
Kelbert, Eugenia 80, 162, 167
Kessels, Geert 81
Kettunen, Kimmo 83
Kilpelä, Eliel 31, 36
Kjus, Audun 19
Korkiakangas, Timo 145
Kosek, Michal 141
Kuoppa, Jenni 48
Kylander, Johan 31
Kyttä, Marketta 48
Lagus, Krista 85
Larsson, Lisbeth 117
Laursen, Ditte 56
Lethbridge, Emily 86
Lindstad, Arne Martinus 27, 29
Lundin, Sverker 87
Lyse, Gunn Inger 68
Misharina, Anna Aleksandrovna 90
Moldovan, Corina Elena 147
Morgan, Paige 21, 23
Movall, Anna 31, 33
Mähler, Roger 90, 139
Mäkelä, Eetu 126
Mårtensson, Lasse 89
Natzén, Christopher 92
Navarretta, Costanza 93
Ní Mháirtín, Sarah Stephanie 96
Nielbo, Kristoffer L. 52
Nielsen, Janne 56
Nielsen, Jonas 101
Nilsson Hammar, Anna 102
Nimb, Sanni 103
Nummi, Jyrki 54
Nøklestad, Anders 141
Nørtoft, Mikkel 106
Offersgaard, Lene 144
Ore, Christian-Emil 7
Ore, Espen S. 108
Paganini, Mario C.D. 110
Paju, Petri 67

Pakkasvirta, Jussi 85
Pálsson, Gísli 111
Pantzar, Mika 85
Parente-Čapková, Viola 111
Pedersen, Gro Benedikte 55
Priestley, Joel 141
Puscasiu, Voica 147
Pääkkönen, Tuula 83
Paavolainen, Maija 126
Rettberg, Scott 112
Risberg, Sara 114
Rockenberger, Annika 73
Ruckenstein, Minna 85
Rudan, Sasha Mile 80, 162, 167
Rudan, Sinisa 167
Rønning, Anne Birgitte 7, 27, 30
Salway, Andrew 115
Saylor, Nicole 16
Schwabe, Fabian 41, 43
Seaton, Mitchell John 121
Sjöberg, Maria 117
Skaltsa, Stella 110
Skare, Erik 120
Skott, Fredrik 19
Skåden, Kristina 41, 44
Snickars, Pelle 139
Solberg, Ragnhild Torvanger 147
Steingrímsson, Steinþór 149
Stenersen, Anne 120
Stolk, Joanne 160
Svensson, Anna 150
Svensson, Patrik 15
Svoboda, Dieta 41p.
Söring, Sibylle 135
Søvsø, Tue Emil Öhler 121
Tahmasebi, Nina 122
Taugbøl, Stine Brenna 151
Tjell, Mette 124
Tolgensbakk, Ida 19
Tolonen, Mikko 126
Tomasek, Kathryn 21p.
Tomasi, Francesca 15
Tveit, Anders Lunde 154
Veit, Elisa 31, 33
Vierros, Marja 155
Wareham, Andrew 21
Westin, Jonathan 47
Westin, Jonathan K. 127
Yli-Siurua, Marjoriikka 85
Zeevaert, Ludger 129

Zhou, Bo 77
Østmoe, Tor Ivar 156
Åhlfeldt, Johan 130

ISBN 978-82-90954-51-7