NarDis

Sauer, Sabrina; Hagedoorn, Berber

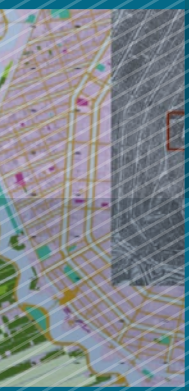
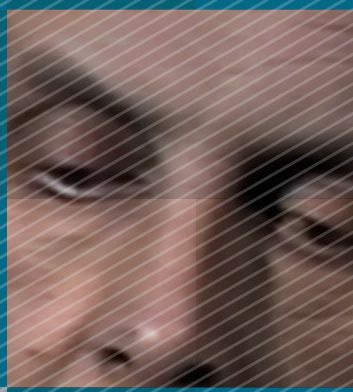
Published in:
CLARIAH

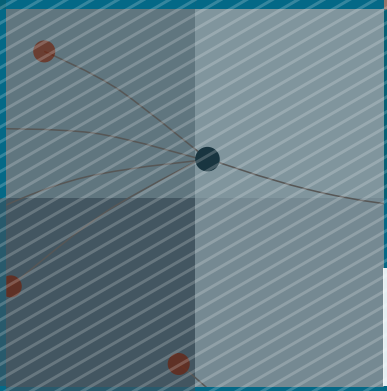
Publication date:
2019



CLARIAH

A digital research infrastructure for humanities researchers



CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities

CLARIAH

A digital research infrastructure for humanities researchers

COLOPHON

Editor: Erica Renckens, Tatataal

Co-editors: Patricia Alkhoven and Arjan van Hessen (CLARIAH)

With thanks to all involved in CLARIAH

Lay out: Linda van den Akker, Akker Ontwerp

Printer: Dpn Rikken Print

March 2019

CONTENTS

WELCOME

CLARIAH-CORE has been a remarkable adventure. The digital turn in our society poses a classic challenge to the humanities: huge opportunities, but numerous hurdles on the way to realisation of the promises of digital scholarship.

In 2009 the CLARIN-NL project initiated the construction of an integrated, distributed digital infrastructure for language-based humanities research, with a focus on linguistics. But the promises of interoperability between different data sets and tools are not restricted to language-based research. Within CLARIN-NL pilot projects had reached out to neighbouring fields. In 2011 a consortium led by Jan Odijk proposed to the 'NWO-Nationale Roadmap voor Grootschalige Wetenschappelijke Onderzoeksinfrastructuur' to build a digital infrastructure for all of the humanities, dubbed CLARIAH. This proposal was rated excellent, but also too ambitious. The consortium was granted seed money, and in 2013, in the next round, we presented a new proposal. As suggested by the vetting committee, the proposal focussed on three disciplines within the humanities where digital techniques already had a large impact: linguistics, socio-economic history and media studies. This choice meant we had to deal with texts, structured data and audio-visual materials. This proposal, CLARIAH-CORE, was funded by NWO, and we started our work from 2015. Over the next few years we built an infrastructure for these fields, but we also cooperated more and more closely between them in building a common humanities infrastructure. When in 2016 we put out a call for pilot studies to test the infrastructure, more than half of the proposals crossed the borders of our disciplines.

We have come a long way together, but we have not yet reached our goal. This has been recognised by our funders, which have granted us the means to continue with the next phase of our project. In CLARIAH-PLUS, which will run from 2019, we will further extend the infrastructure. We will now be able to include analysing texts for fields such as literature, history or philosophy, which look more to their contents than to the language they are in (even if that border is permeable too). We will doubtlessly cross borders to other disciplines within and outside the humanities.

But first we invite you to celebrate with us the accomplishments of CLARIAH-CORE in this booklet. We wish you an enjoyable read.

Lex Heerma van Voss, PI CLARIAH
Jan Odijk, director CLARIAH

INTRODUCTION

CLARIAH (*Common Lab Infrastructure for the Arts and Humanities*) is a digital research infrastructure for humanities researchers in the Netherlands. It forms an integrated part of the European CLARIN and DARIAH research infrastructures. The design, construction and exploitation of the CLARIAH infrastructure is being carried out in a series of projects that will be described in this booklet. The longer term sustainability of CLARIAH (beyond the limited lifespan of the projects) is ensured by a network of data and service centres firmly embedded in humanities research.

Before elaborating on this, we first sketch an important development in the humanities that made CLARIAH necessary and possible.

THE DIGITAL TURN

The amount of data that is available in digital form is increasing exponentially. This is true generally, but also for the humanities in the Netherlands. It includes contemporary newspapers, journals, TV and radio broadcasts (texts of 1.5 million radio bulletins), new media (twitter, Facebook, etc.), but also historical newspapers (over 80 million articles in Delpher), books (over 170 k retro-digitized books from the 18th- 20th century and e- books) and magazines (over 1.5 million pages from the 18th and 19th century), digitized and born digital archival materials, structured data sets, etc.

The fact that the data are becoming available in digital form implies that they can be analysed with digital techniques. In addition, the computer hardware enables this processing, basic analysis software is available, and advanced analysis techniques, inter alia natural language processing tools and applications, often yield sufficient quality to use them. Therefore, the so-called *Digital Turn* creates huge opportunities for the humanities: it can broaden the empirical basis for its research, since digital techniques can analyse data in quantities that humans never could cover. It will therefore enable the investigation of existing research questions in new ways, create opportunities for investigating research questions that could not be addressed before, and for formulating and investigating completely new research questions.

However, this digital turn is not going to be easy! The reason is that this enterprise involves the entire spectrum of the analytics challenge of big data: we have to work with massively distributed data sources, both structured data and unstructured data, of varying complexity and quality (often with a lot of noise, partially incomplete, etc.). Large volumes of unstructured data come in multiple formats: audio, video, image, text, requiring formal (syntactic) and semantic interoperability. And the users of these digital data are globally distributed, and highly varied, across many humanities disciplines, all speaking very different languages and working in different research traditions. In addition, there are big differences among the humanities researchers in terms of their technical knowledge and expertise, and their willingness to embrace the digital techniques.

These problems are also familiar to modern software companies active in the area of text analytics, which attempt to analyse noisy digital texts and their metadata in a wide variety of formats and from a wide variety of social media platforms. These include language technology companies, but also their customers, and they include big players such as Google, Facebook, and IBM. IBM explicitly recognizes the parallels: “The challenges faced by the Art & Humanities are highly representative and synergistic with

the broader challenges IBM is solving across other industries – from law enforcement to health care and beyond”. In addition, the problem is familiar to public digital heritage organisations.

These problems make it necessary to set up a research infrastructure for the humanities to facilitate the Digital Turn: CLARIAH.

RESEARCH INFRASTRUCTURE

CLARIAH is a digital research infrastructure for humanities researchers. An *infrastructure* is a set of usually large-scale basic physical and organizational resources, structures and services needed for the operation of a society or enterprise. Typical examples are the railway network, the electricity

CREDITS: JØRGEN STAMP (CC BY 2.5 DK)

network, or on a smaller scale, the availability of wireless internet through Eduroam at all Europe's educational premises. A *research infrastructure* is an infrastructure intended for carrying out research: facilities, resources and related services used by the scientific community to conduct cutting-edge research. Famous examples are the Chile Large Telescope and the CERN Large Hadron Collider. CLARIAH is a *digital* research infrastructure because it focuses on digital data. *Humanities researchers* include linguists, historians, literary scholars, philosophers, religion scholars, and others, and include (in the CLARIAH context) researchers that usually are counted as social scientists, in particular political sciences researchers.

The CLARIAH infrastructure is distributed (its data and services run on servers of multiple centres) and virtual, i.e. only accessible digitally via the internet.

INTERNATIONAL CONTEXT

CLARIAH forms an integrated part of the European CLARIN and DARIAH research infrastructures.

CLARIN (Common Language Resources and Technology Infrastructure) focuses on *language* and therefore provides facilities for digital language resources. Digital language resources include software and data. The data include textual data in natural language, databases about natural language (typological databases, lexical databases, dialect databases, etc.), and audio-visual data containing (written, spoken, signed) language. The software includes programmes for analysing language in textual and audio-visual data, for enriching language data with a wide variety of linguistic annotations, and for searching in language data that contain these linguistic annotations. CLARIN considers language in all the functions it is used for, not only as an object of inquiry, but also as a carrier of cultural content, as a means of communication, and as a component of identity. CLARIN has set up an ERIC, a legal entity at the European level specifically set up for research infrastructures, which is hosted by the Netherlands. CLARIN ERIC currently has 20 member countries, four observer countries, and a cooperation agreement with one party from outside of the EU.

DARIAH (Digital Research Infrastructure for the Arts and the Humanities) aims to enhance and support digitally-enabled research and teaching across the Humanities and Arts. It is a network of people, expertise, information, knowledge content, methods, tools and technologies coming from various countries. DARIAH also set up an ERIC, which is hosted by France, and currently has 17 member countries and cooperating partners in 8 countries.

CLARIN and DARIAH are both distributed infrastructures. CLARIN is implemented in a network of CLARIN centres. These centres come in different flavours and include centres for general infrastructure services, centres for data and software services, and (virtual) knowledge centres.

DARIAH is organized in Virtual Competence Centres (VCCs), of which there currently are four. Each VCC focuses on a particular theme: e-infrastructure, the liaison between research and education, the management of scholarly content, and advocacy, impact and outreach.

CLARIN and DARIAH are also both *distributed* and *virtual* infrastructures: most of their organizational units are implemented in a distributed, international and often cross-disciplinary network of actual organisations, and both infrastructures provide their services mainly via the internet.

NATIONAL ROADMAP PROJECTS

The Netherlands maintains a national roadmap for large scale research infrastructures. The first humanities project on the Netherlands' large scale research infrastructure roadmap that was awarded funding was the CLARIN-NL project (2009-2015), which exclusively focused on CLARIN. In 2011, CLARIN and DARIAH in the Netherlands decided to join forces. This resulted in the CLARIAH-SEED project (2012-2014). The main project described in this booklet is the CLARIAH-CORE project (2015-2019). It will be succeeded by the CLARIAH-PLUS (2019-2023) project.

CLARIAH-CORE

The humanities cover a wide range of disciplines, data types, approaches, methodologies and traditions. Creating a single digital research infrastructure for the whole humanities is quite a challenge and there is a serious danger of losing focus if all the different disciplines have to be accommodated at once. In order to avoid this risk, CLARIAH-CORE focuses on three disciplines within the humanities:

- Linguistics
- Social-economic History
- Media studies

And on the main data types used by these disciplines:

- Natural language text
- Structured (often quantitative) data
- Audio-visual data

Since these data types are covered by the core disciplines, and since many data and tools for these core disciplines are also relevant for other humanities disciplines, we expect that the limitation to these three core disciplines will not impede later extensions to other disciplines within the humanities. The selection for these three disciplines was also motivated by the fact that they are forerunners in Digital Humanities in the Netherlands.

Almost all universities in the Netherlands with a humanities department, as well as Royal Academy research institutes and independent research institutes participate in the CLARIAH-CORE project. The network of centres that ensures the sustainability of the infrastructure consists of research institutes that also provide data and software services (Huygens ING, IISH, Meertens Institute, Dutch Language Institute) as well as dedicated data and/or software service centres (DANS, NISV).

CLARIAH-CORE provides **generic** infrastructure services and data: facilities for shared vocabularies, for (meta)data as Linked Data, and for search in the linked data (ANANSI), access control, CMDI to linked data conversion and vice-versa, an OCR/Text Correction and enrichment pipeline (PICCL), guidelines for standardization, and facilities for performance and availability of services.

In the **linguistics** work package, the goal is to support the linguist in each stage of a research project. For each of these stages it has been inventoried what was needed and what was already available from earlier projects. Additions to and extensions of the existing functionality have been defined and implemented. A (as yet incomplete) overview of resulting tools and services implemented in a faceted search interface can be found here: http://bit.ly/CLARIAHtools. Metadata for most data have been incorporated in the CLARIN Virtual Language Observatory, and a new curated metadata catalogue of Dutch language corpora has been made available here: http://bit.ly/CollectionBank.

In the **socio-economic history** work package databases at the *macro* (national/international), *meso* (trade unions, organisations) and micro (individual / family) levels are being linked. These databases have different histories, are structured in incompatible ways, and use different vocabularies. Integration of these databases is carried out using the Linked Data paradigm, and this integration will enable addressing research questions that require relating social-economic facts from different levels.

In the **media studies** work package the researchers will be supported by integrating improved versions of a range of independently developed applications in one virtual research environment *Media Suite*.

The applications include CoMerDA, an aggregated search interface for audio-visual data; AVResearcherXL for exploring audio-visual metadata in historical context; TrOve (Transmedia Observatory), a search application to analyse the distribution of information throughout time across different media; DIVE, presentation of collection items in context and 'intuitive' browsing, and OHT (Oral History Today), which supports the full workflow of working with unstructured audio-visual content.

RESEARCH PILOTS

Research pilot projects are small research projects aimed at testing the infrastructure and/or specific parts of it. Such projects will lead to improved functionality, driven by concrete needs of humanities researchers working on one or a few closely related very concrete research questions.

ADAH PROJECTS

CLARIAH and the NL eScience Center together set up a call for humanities projects that stimulate and illustrate the acceleration and upscaling of humanities research that can be achieved by applying advanced ICT methods to humanities data and research problems. Four projects were awarded funding. These projects are still running and will finish in the course of 2019.

INTEGRATION PROJECTS

In order to demonstrate the potential of the CLARIAH infrastructure for cross-disciplinary work a number of projects were started up.

In the **Athena project** different data types (textual sources, structured data, and audio-visual material) from the biodiversity heritage domain are combined on one platform and integrated with CLARIAH.

The **Amsterdam Time Machine** (ATM) aims to use historical Linked Open Data (LOD) on Amsterdam to create a web of data on people, places, relations, events and objects and present them within their own context in terms of time and place using geographical and 3D-visualisations.

DIGIFIL aims to digitise the Dutch Filmladders (the weekly listings of movie showtimes at local cinema theatres or other venues) and contextual information about the wider movie landscape as reported in historical newspapers (such as movie reviews and descriptions).
It integrates the disciplines linguistics and media studies, and textual, structural and audiovisual data types.

CLARIAH-PLUS

The CLARIAH-CORE project is in its last year. A successor project, called CLARIAH-PLUS (2019-2023), has already started. In CLARIAH-PLUS we extend the scope of the research infrastructure to the treatment of text as a carrier of content. This is needed for many humanities disciplines, e.g. for literary studies, history, religion studies, and philosophy.
The National Library, with its huge amount of digitised textual material and facilities to search in these data, has joined CLARIAH-PLUS and plays a crucial role as a centre for textual data, not only for providing data, but also for offering services for processing data. The latter is needed to avoid a proliferation of different versions of data at multiple locations and to ensure a proper handling of IPR and other legal restrictions.

The following pages give an overview of where CLARIAH stands at the end of CLARIAH-CORE. What we hope will jump from the pages is that we have worked enthusiastically at building an infrastructure for Digital Humanities. We strongly believe that achieving interoperability of data and tools will open up a treasure trove of research possibilities. We also are convinced that we are not there yet. We will as enthusiastically pursue this goal in CLARIAH-PLUS, even if we suspect that there will still be further work to do at the end of that phase of CLARIAH.

DISSEMINATION

The Dissemination and Outreach team is responsible for the organisation of the general meetings, the website, education, and communication (newsletters, tweets) with and about the CLARIAH community.

TOOGDAYS

The Dissemination Team organises annual CLARIAH 'Toogday', a general meeting with presentations and demonstrations of ongoing projects for all involved or interested in CLARIAH. Five Toogdays were held: a kickoff-meeting in 2015, one in 2016 and 2017 and in 2018 two well-attended Toogdays took place (9 March with 83 participants and 19 October with about 70 participants). The second Toogday was meant specifically for presentations by the 16 Research Pilots. The Toog days were animated by demonstrations on big screens during the breaks and drinks.

TECHDAYS

Techdays are technical working sessions across the work-packages for CLARIAH (related) technicians. They are mainly meant for the real developers and less suitable for the more general CLARIAH-audience. Therefore they have a relatively small-scale character. Four Techdays were held since 2015; most were visited by about 30-40 people. On an average Techday, the participants work together trying to solve each other's problems.

CLARIAH ON TOUR

During the first years, the concept of an 'infrastructure for the humanities' was little or not known to most humanities scholars. If scholars knew CLARIAH, it was mostly associated with 'some research programme where you can get money'. Moreover, it was noticeable that most of the responses to CLARIAH calls came from those institutions that were somehow actively involved from the beginning of CLARIAH. In order to include the less active humanity faculties, CLARIAH on Tour was set up to inform humanities and social sciences scholars about the possibilities of the new infrastructure (with new tools and Big Data technologies) for their research.

CLARIAH on Tour visited the University of Groningen, Leiden University and Utrecht University. In each event about 40-50 participants took part. An evaluation in Utrecht (2018) showed that although most participants appreciated the more general overview given, some of the participants had wanted to hear more technically specific matters and/or substantive matters related to CLARIAH. Because the story presented was relatively generic, very specific questions could not always be answered.

After 4 years of CLARIAH, visits to faculties, participation on events and of course the Toog- and Tech days, we may conclude that CLARIAH is known by a large part of the humanity scholars. Therefore, the next CLARIAH on Tour events will be organised differently. Instead of explaining the 'what and why' of CLARIAH, we will focus on the needs of the groups to visit. For example: if a group of humanity scholars is heavily working on AV-media, the CLARIAH on Tour visit will be done by representatives of WP5 in order to present more specific cases to the audience and answer their AV-related questions.

CLARIAH COURSE TASK FORCE

The purpose of the CLARIAH Course Task Force was to bring together teachers at Dutch universities of courses closely related to Digital Humanity and let them share knowledge about their DH-courses. By telling and showing these teachers about the tools and data available in the CLARIAH infrastructure we hoped to let them act as ambassadors for CLARIAH. Moreover, the assembled teachers generated ideas for setting up a national teaching platform for Digital Humanities and integrate CLARIAH modules in their universities curriculum.

The intention to provide Digital Humanities teachers with a platform for knowledge exchange and to share information what each university is doing in this area and what their specific approach is, has worked well. Further, the need appeared act together to tackle generic DH education issues - e.g. how do we implement CLARIAH in the various curricula. Prior to that, this platform was still lacking in national mandate and executive power. An interesting spin-off is the education platform Ranke.2 (https://ranke2.uni.lu/).

WEBSITE

The CLARIAH website is the main dissemination platform. It is used for all communication about CLARIAH (related) (inter)national events, CLARIAH calls, courses, summer schools, blogs about events interesting for the CLARIAH community, videos, PowerPoint presentations and more. In addition, it allows for the submission of requests for travel and subsistence expenses. The lay-out of the website is such that information is easily accessible for various CLARIAH group of stakeholders, from the researchers and technicians to the International Advisory Board.

NEWSLETTER

The CLARIAH newsletter has been published four times a year with news reports about CLARIAH and CLARIAH related events and projects. It was send to a mailing list with ca 375 subscriptions. The newsletter is partly filled with CLARIAH specific items, partly with information from our European sister organisations DARIAH-EU and CLARIN ERIC.

SOCIAL MEDIA

Social media such as Twitter and Facebook are seen as ‘the’ media for keeping the target group informed about CLARIAH’s activities. With Twitter this works reasonably well, but with Facebook we stopped due to the absence of interest. Apparently the CLARIAH target group can’t be found (anymore) on Facebook.

The @CLARIAH_NL tweets are shown on the website, as are those of ‘sister organisations’ such as @CLARINERIC and @DARIAH-EU and @Parthenos_EU.

VIDEOS

There are five short videos about CLARIAH general and work packages and work package-transcending topics. The films can be watched via youtube and the CLARIAH website.

E-DATA & RESEARCH

e-Data & Research is the magazine that distributes news about e-research projects and CLARIAH-related subjects. The magazine appears three times a year and is freely distributed to all social science and humanities researchers at Dutch universities and research institutions. CLARIAH is contributing by providing the editor of e-Data & Research.

CLARIAH PARTNERS

RESEARCH PILOTS

Research pilot projects are small research projects aimed at testing the infrastructure or specific parts of it. Research pilots therefore entail the cross-domain cooperation of the groups and institutes that have built or that make available relevant parts of the infrastructure. Such projects will lead to improved functionality, driven by concrete needs of humanities researchers working on one or a few closely related very concrete research questions. It may lead to successfully concluded research, new requirements for the infrastructure or particular applications, services or data within the infrastructure.

2TBI

TOWARDS AN INTERNATIONAL BIOGRAPHICAL INFRASTRUCTURE

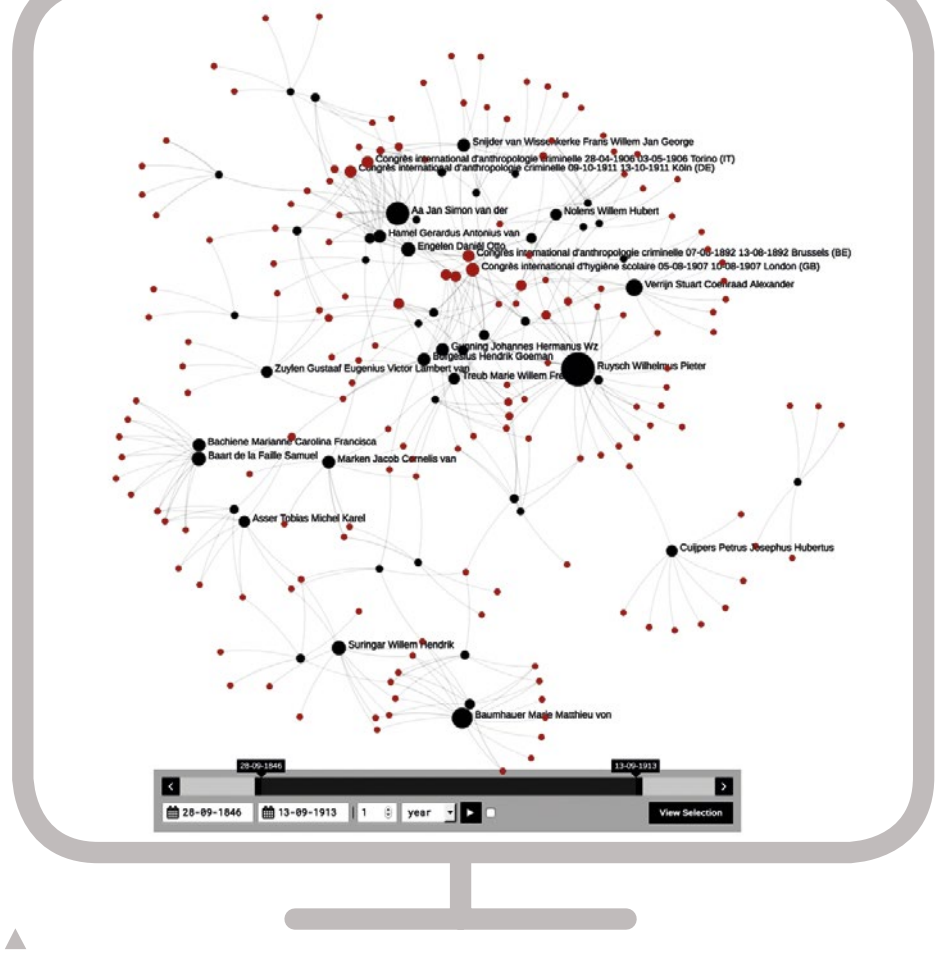
MAASTRICHT UNIVERSITY,
GHENT UNIVERSITY,
HUYGENS ING AND LAB1100

CONTACT:
NICO RANDERAAD
N.RANDERAAD@MAASTRICHTUNIVERSITY.NL

The 2TBI-team set out to link a database of persons who were internationally active in the 19th and early 20th century, with online biographical resources in the Netherlands. To put it in plain terms, we wanted to know more about the local and national backgrounds of Dutch reformers who – we know – were involved in initiatives at an international level. The result of our endeavor is a selection of around 1100 Dutch persons, whom we can trace in various data collections (see the dataset on the Clariah infrastructure).

2TBI was important in gaining experience with the ResourceSync protocol for harvesting data. At the end of the pilot, the set-up of the ResourceSync connection between the Nodegoat software, used by the researchers (cf. the parent project TIC based in Ghent), and Anansi was running successfully.

Apart from the technical advances, the project also pursued promising research lines in transnational history. 2TBI's research objective is to show to what extent and in which ways Dutch social reformers were active at the local level, on a national



DUTCH PARTICIPANTS (BLACK NODES) WHO ATTENDED MORE THAN 5 INTERNATIONAL CONGRESSES (BROWN NODES), VISUALIZED IN NODEGOAT

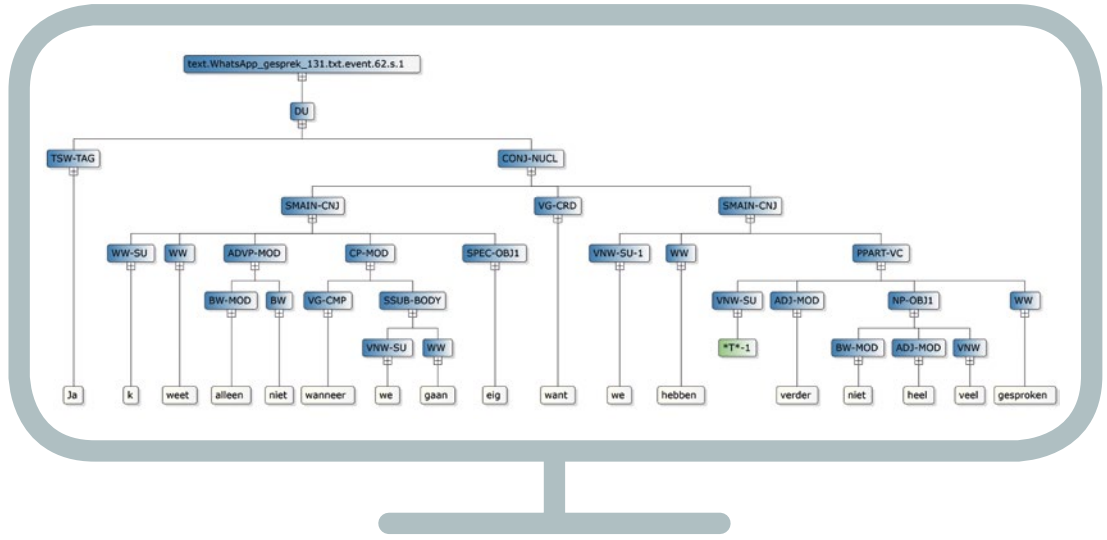
scale, and at international congresses, in order to explore the transnational embeddedness of the reform issues in which they were involved. Our (ongoing) research not only looks into the organizations which the reformers represented or were affiliated with, and which were mentioned in the congress proceedings, but also probes further into local and national backgrounds that emerge from other sources (national and/or specialized biographies, almanacs, address books, library catalogues digital resources, etc.). To our surprise, the names of quite a few internationally active reformers do not appear in standard national biographies.

ACAD

AUTOMATIC COHERENCE ANALYSIS OF DUTCH

RADBOUD UNIVERSITY NIJMEGEN,
UTRECHT UNIVERSITY
CESAR.SCIENCE.RU.NL

CONTACT:
WILBERT SPOOREN,
W.SPOOREN@LET.RU.NL

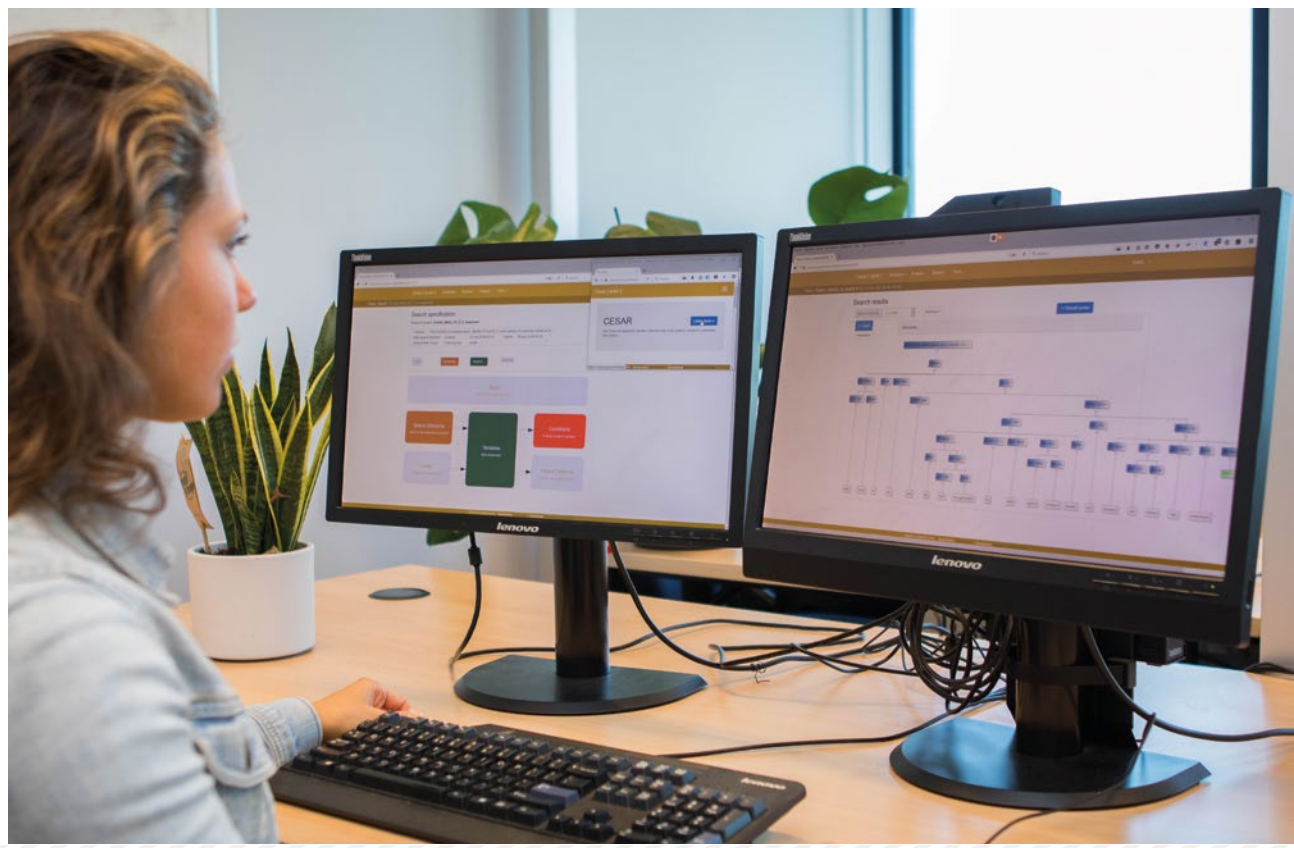


GRAMMATICALLY PARSED TEXT MESSAGE WITH THE DUTCH COHERENCE MARKER 'WANT'.

The goal of ACAD was to develop an environment in which computationally naive discourse analysts can carry out an automatic analysis of causal coherence in discourse.

The research question of the project is: To what extent do the results of small-scale causal coherence analyses in different genres in terms of subjectivity hold for large datasets?

Coherence markers such as *want* and *omdat* differ in their degree of subjectivity. As a discourse analyst, one wants to be able to investigate the environment of these markers, to see whether the environment of subjective connectives like *want* contains more subjective words than that of relatively objective connectives like *omdat* and *doordat*. The ACAD tool allows the researcher to search through a large number of corpora (some already available in CLARIAH, like SoNaR, some newly added, like a corpus of Dutch WhatsApp messages). Core of the project is a search interface, CESAR (Corpus Editor for Syntactically Annotated Resources). CESAR allows the user to formulate advanced search queries without any advanced programming skills. It makes use of the annotations available in the corpora (POS-tagging, lemmatization, grammatical parse). It also has many options to control the output. In principle, the search interface is extendable to other languages and other types of research questions.



CoDoSiS

COMBINING DATA ON SLAVERY IN SURINAM

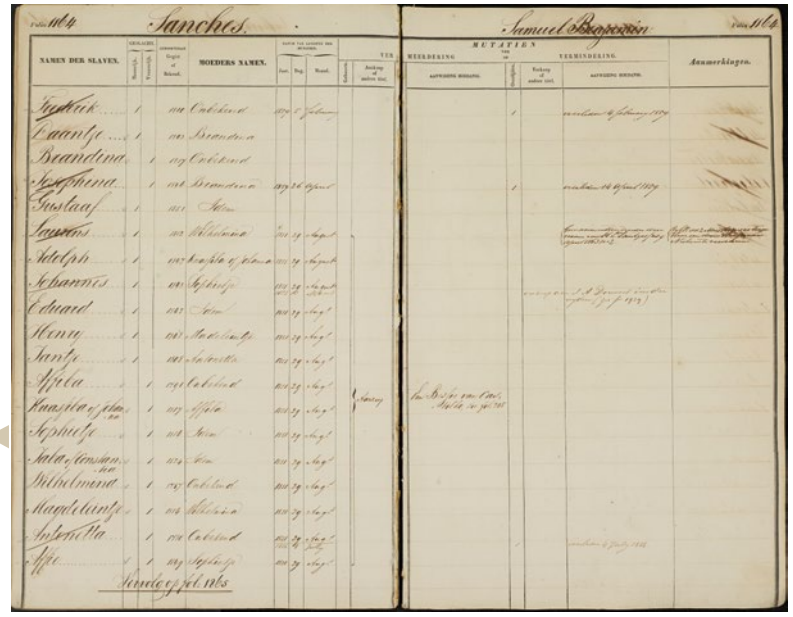
RADBOUD UNIVERSITY,
INTERNATIONAL INSTITUTE OF SOCIAL HISTORY

CONTACT:
COEN VAN GALEN
C.VANGALEN@LET.RU.NL

A SLAVE MARKET IN SURINAM (BENOIT, VOYAGE À SURINAME 1839).

The aim of the research pilot 'Combining Data on Slavery in Surinam' (CoDoSiS) was to develop a strategy to convert existing datasets on slavery in Surinam into Linked Data and to combine them into one database network with relevant connections. The need for this research pilot project arose because in the past two decades a number of datasets and digital transcriptions have been made of sources related to slavery in Surinam in the 19th century. Many of these sources used different file formats and different structures, which makes it hard to combine them in one meaningful database network.

CoDoSiS did not aim to create a new database, but instead opted for a construction to link existing datasets, by converting them into Linked Data by using the CLARIAH wp4-tool QBer and to combine them into one database network with relevant connections using the CLARIN-tool TICCL. During the process, we were able to replace QBer by a new and more effective CLARIAH tool, CoW. We created a pilot based on four different datasets in which we showed the feasibility of the proposal.



A FOLIO OF THE SLAVE REGISTER OF SURINAME (NATIONAL ARCHIVE OF SURINAME, INVNR 26, FOLIO NR. 1164).

CrossEWT

CROSS-MEDIAL ANALYSIS OF WW2 EYEWITNESS TESTIMONIES

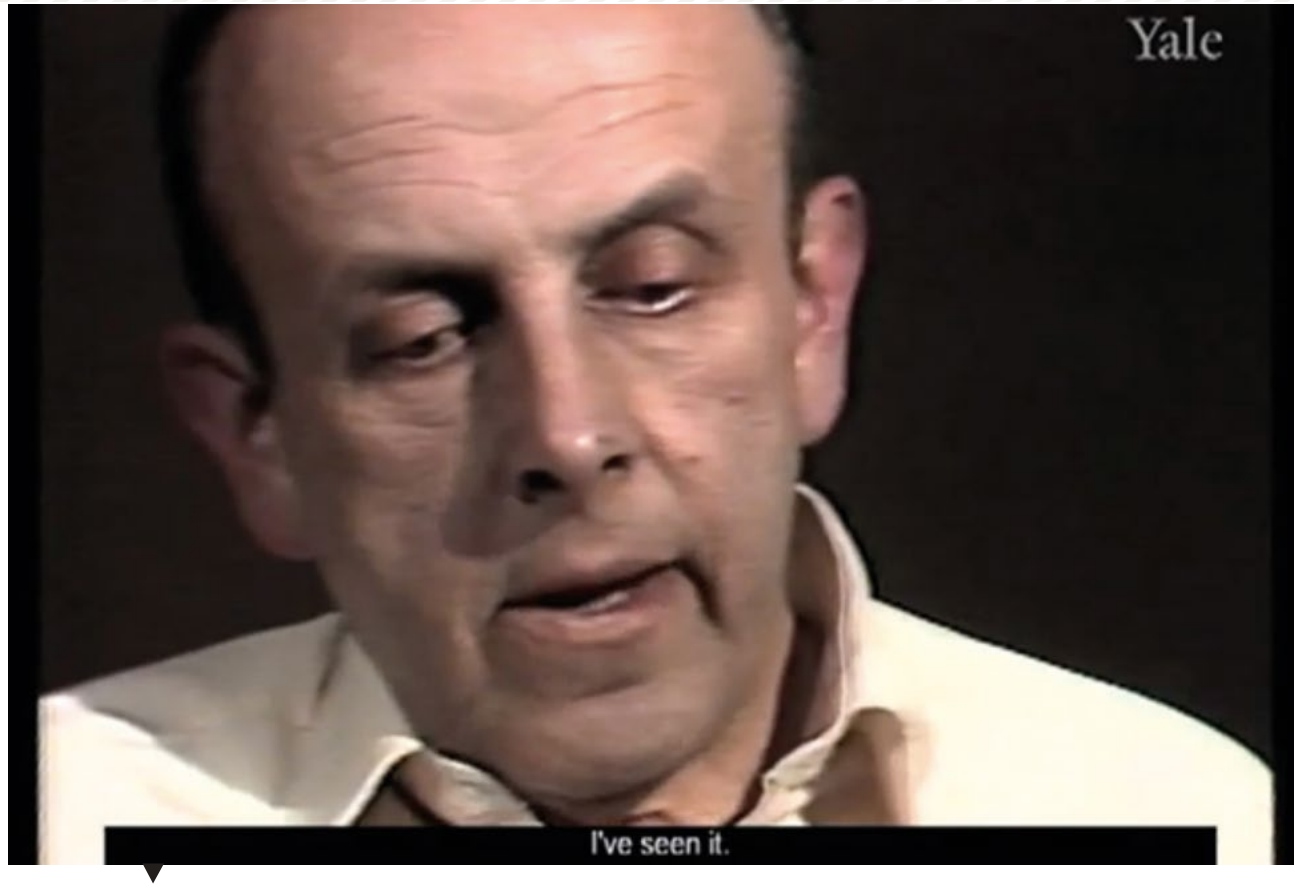
ERASMUS UNIVERSITY ROTTERDAM, NETHERLANDS
INSTITUTE FOR SOUND AND VISION, DANS,
NETWERK OORLOGRBRONNEN

CONTACT:
SUSAN HOGERVORST
SUSAN.HOGERVORST@OU.NL

Since the 1960s, eyewitnesses have become ever more mediatized, and ever more prominent in popular representations of the Second World War. Many initiatives have been undertaken to preserve their accounts. One of the most large-scale examples is the Visual History Archive, containing over 52,000 video-interviews about the Shoah. In the Netherlands, hundreds of WW2-related oral history interviews are filed at DANS.

Despite the prominence and mediatization of eyewitnesses, there is no systematic research about which topics have actually been addressed in their accounts. This project entails a diachronic content analysis and comparison of eyewitness testimonies (EWTs) about the Second World War in the Netherlands. The focus is on testimonies that have been published since 1945 and that have been generated in three different, but interrelated media contexts: newspaper articles, television documentaries, and oral history interviews. The data therefore consists of newspaper articles, transcriptions of documentaries generated with automatic speech recognition, and interview transcripts of the open access 'Getuigenverhalen' interview collection as hosted by DANS.

In the Clariah Media Suite, relevant collections that contain EWTs have been inspected. Thereafter, three subcorpora have been created and exported to text analysis tools with which their content could be analyzed and compared systematically.



STILL FROM LEON S. EDITED TESTIMONY.(FORTUNOFF ARCHIVE, YALE UNIVERSITY)

A USC STUDENT LISTENS TO A TESTIMONY IN THE SHOAH FOUNDATION'S VISUAL HISTORY ARCHIVE. THE ARCHIVE CONTAINS 52,000 TESTIMONIES FROM SURVIVORS OF THE HOLOCAUST AND OTHER GENOCIDES.

DB:CCC

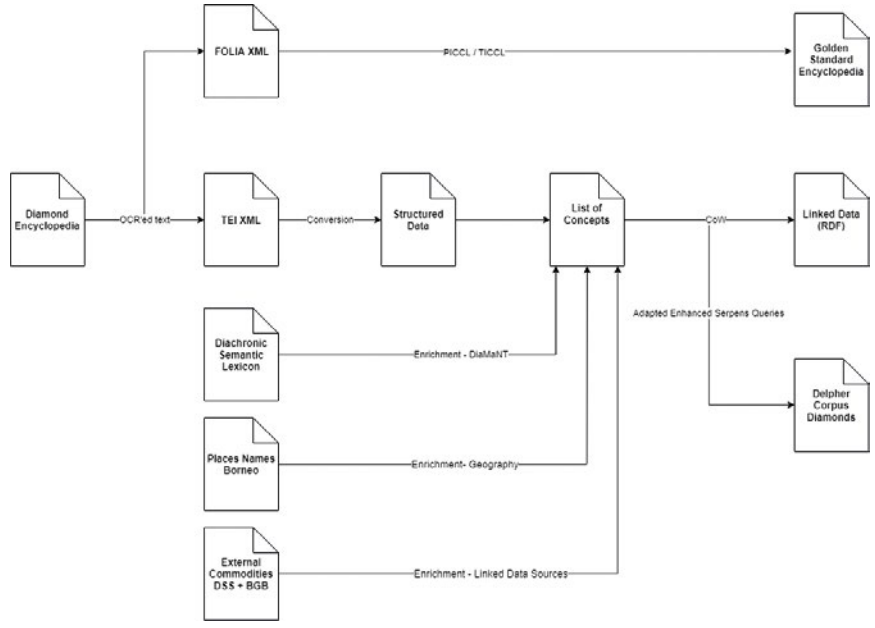
DIAMONDS IN BORNEO: COMMODITIES AS CONCEPTS IN CONTEXT

INTERNATIONAL INSTITUTE OF SOCIAL HISTORY,
TILBURG UNIVERSITY,
DUTCH LANGUAGE INSTITUTE, VU AMSTERDAM

CONTACT:
KARIN HOFMEESTER
KHO@IISG.NL

The intensified circulation of people, commodities and ideas is one of the characteristics of a globalizing world. If we want to understand the causes and consequences of these circulations, we have to know which commodities circulated when and where and who circulate them.

Our project uses diamonds as a pilot, more specifically diamonds in Borneo, so far a true blind spot in our knowledge on the global diamond commodity chain. We know little on where diamonds were found, who the miners and traders were and if there was really an 'age-old' diamond polishing industry as is sometimes suggested. To answer these questions we developed a workflow that enables us to query the journal corpus of Delpher in an efficient and elaborate way that can also be used for research on other commodities. Following this workflow, we used the 1908 *Geillustreerde encyclopaedie der diamantnijverheid* as a starting point for our concept list. The text is converted into structured data and a Gold Standard version of the text is created with TICCL. The concept list is enriched with synonyms and historical variant spellings (DiaMaNT); historical place names on Borneo and external Linked Data Sources. Adapted versions of enhanced scripts made for the CLARIAH project Serpens are used to query the journal corpus of Delpher.



THE DB:CCC WORKFLOW.

DIAMANTMIJN. UIT: SCHWANER, C.A.L.M (1853), BESCHRIJVING VAN HET STROOMGEBIED VAN DEN BARITO EN REIZEN LANGS EENIGE VOORNAME RIVIEREN VAN HET ZUID-OOSTELIJK GEDEELTE VAN DAT EILAND, VOLUME 1., P.N. VAN KAMPEN, AMSTERDAM.

DReAM

DEBATE RESEARCH ACROSS MEDIA

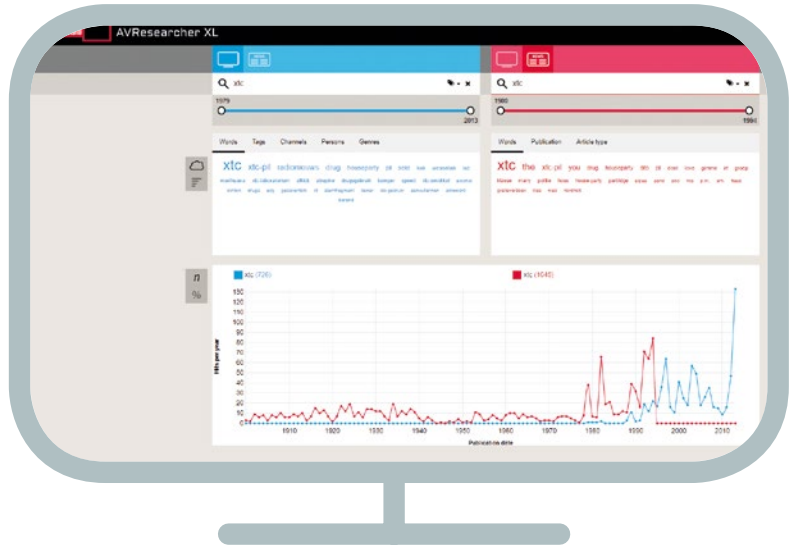
CROSS MEDIA RESEARCH OF PUBLIC DEBATES ON DRUGS AND REGULATION

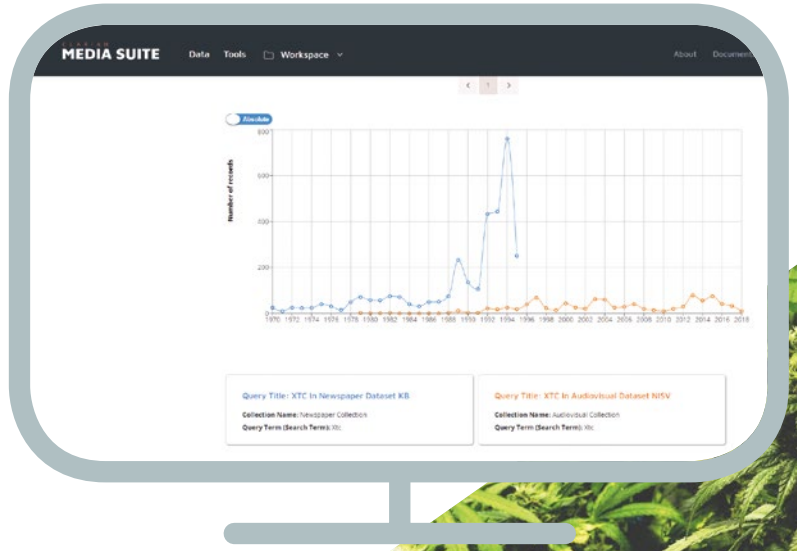
UTRECHT UNIVERSITY,
NATIONAL LIBRARY OF THE NETHERLANDS,
NETHERLANDS INSTITUTE FOR SOUND AND VISION

CONTACT:
TOINE PIETERS
T.PIETERS@UU.NL

In this research pilot historians have tested and extended CLARIAH's tool AVResearcherXL.

We are in the process of studying the role of historical public debates on drugs and regulation (1945-1990) in newspapers, on radio and television. These debates are shifting in time and often fragmented since drugs (e.g. heroin, amphetamines and cannabis) move between medical, criminal and recreational spheres. We study the historically dynamic relation between governmental drug regulation and public discourse. To do that, we aim to enable our research strategy, which is to trace and understand public debates by alternating between distant reading and close reading, across textual and AV-datasets. AVResearcherXL is primarily developed as a distant reading tool with a focus on media representation research. By enriching the AVResearcherXL tool with additional CLARIAH components we have made it suitable for alternating cross-media forms of distant and close reading. This significantly improves the employability of AVResearcherXL for humanities researchers.





SCREENSHOTS FROM THE MEDIA SUITE WITH QUERIES ON DRUGS AND REGULATION.

HUMIGEC

HUMAN CAPITAL, IMMIGRATION AND THE EARLY MODERN DUTCH ECONOMY

HUYGENS ING,
INTERNATIONAL INSTITUTE
FOR SOCIAL HISTORY

CONTACT:
JELLE VAN LOTTUM
JELLE.VAN.LOTTUM@HUYGENS.KNAW.NL

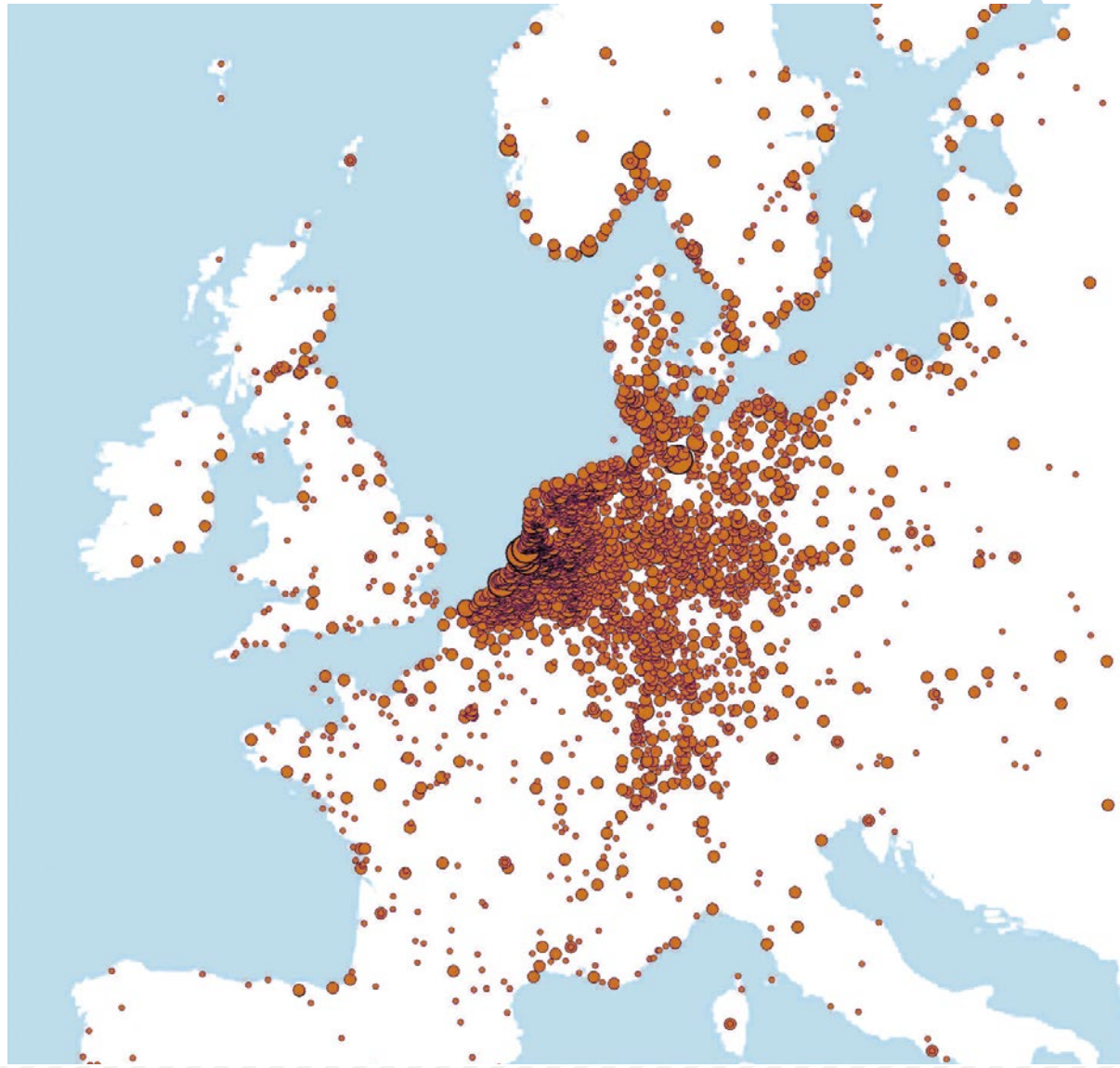
Job mobility of native and immigrant workers in the maritime labour market, c.1700-1800

What was the contribution of migrant workers to the 18th-century Dutch economy? We reconstructed the careers of native and migrant sailors who worked for the Dutch East India Company (VOC) and analysed these to gain insight into the skill levels of migrant workers.

We took an existing dataset consisting of 800,000 employment records as a starting point, standardised the workers' birthplaces and developed an automated entity matching tool that allowed us to reconstruct individual careers. The tool first normalises spelling variations and then generates clusters on the basis of name similarity and a set of date conditions.

Our research findings will be published in a forthcoming paper, but our initial analysis of the careers dataset provided some very interesting results. While at the beginning of the 18th century native workers tended to have more successful careers, by the end of the century migrant workers were promoted more often (and were therefore more successful) than their domestic counterparts. This is an important conclusion and one we could not have reached without the tool we developed in HUMIGEC.

VISUALISATION OF THE BIRTHPLACES OF THE 800,000 WORKERS IN THE DATASET USED IN HUMIGEC.



LinkSyr
LINKING SYRIAC DATA

VU AMSTERDAM,
DANS

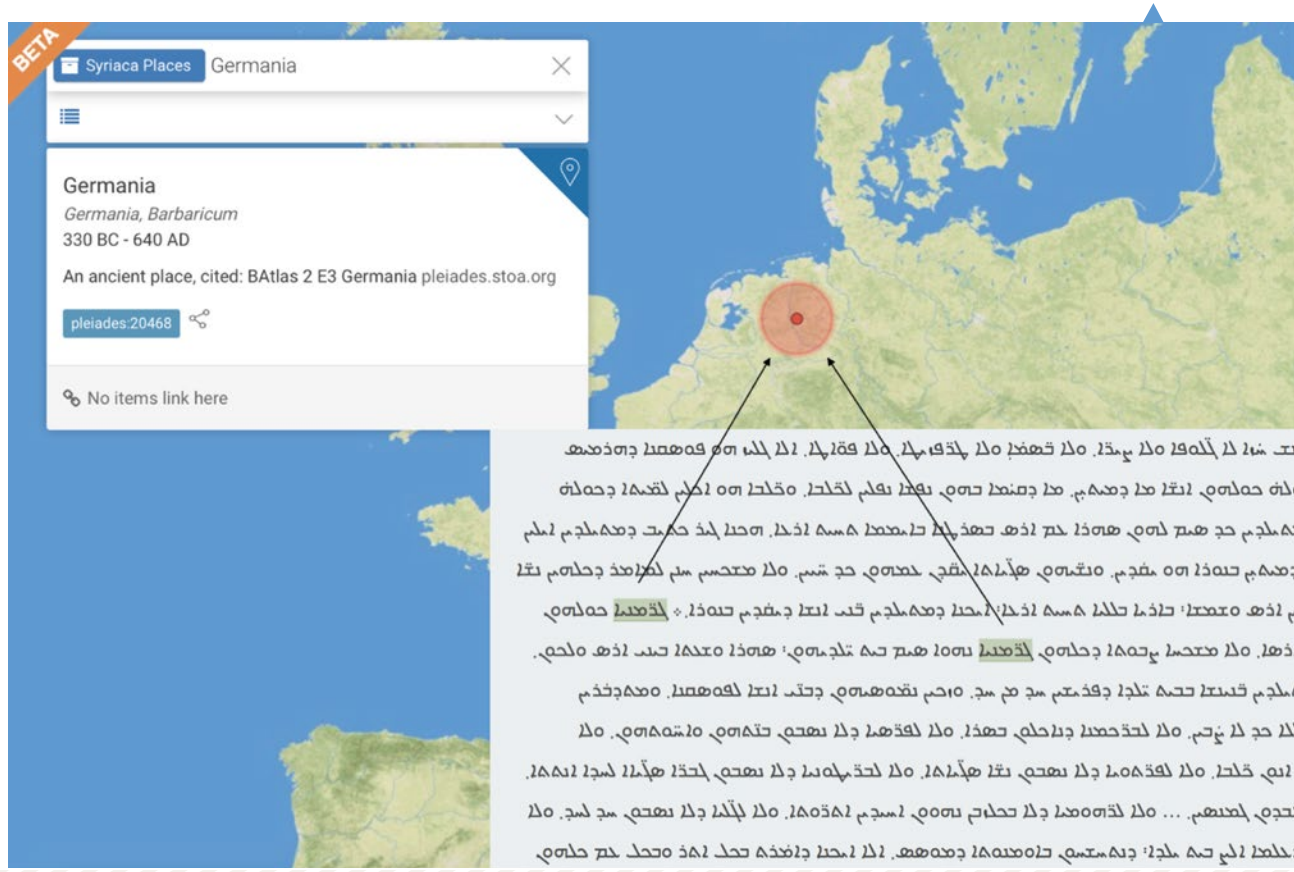
CONTACT:
WIDO VAN PEURSEN
W.T.VAN.PEURSEN@VU.NL

How do the Biblical heritage and Hellenistic culture interact in the oldest documents of Syriac Christianity?

The Eep Talstra Centre for Bible and Computer (ETCBC) investigated this question in the CLARIAH pilot project LinkSyr (2017–2018), using linguistic data processing, especially topic modelling. The Syriac Book of the Laws of the Countries (BLC), written by the 2nd/3rd-cent. Syriac philosopher Bardaisan is compared with the ancient Syriac translation of the Bible ("Peshitta", 2nd cent.), other ancient sources. The analysed texts are exposed as Linked Open Data and related to the lexicographical and encyclopedic resources of Syriaca and SEDRA. The former presents the URIs for a large number of place names and person names for the Syriac heritage, whereas the latter contains dictionary information for a list of more than 50,000 lexemes. We developed a pipeline for the analysis of Syriac texts from OCR through data preparation, parsing and NER to Linked Data. Thanks to a Pelagios Research Development grant we could link this project to the Pelagios infrastructure. A grant ('KDP') from DANS enabled us to collect also Syriac liturgical data from the collection of the Peshitta Institute. The project included a workshop on Linked Data and Syriac Sources (March 2018) and a bootcamp on NLP tools for Syriac (January 2019).

The project results of LinkSyr are stored in a github repository. In the near future the data will also be presented through syriac.ancient-data.org as well as peshitta.ancient-data.org (data of the ancient Syriac Bible translation used in the project) and lectionaries.ancient-data.org (structured liturgical in related to the textual data).

A TOOL DEVELOPED IN THE LINKSYR AND PELAGIOS PROJECTS ENABLES THE CONNECTION BETWEEN NAMED ENTITIES IN ANY SYRIAC TEXT AND HISTORICAL MAPS OF THE PELAGIOS INFRASTRUCTURE.



M&M

ME & MYSELF

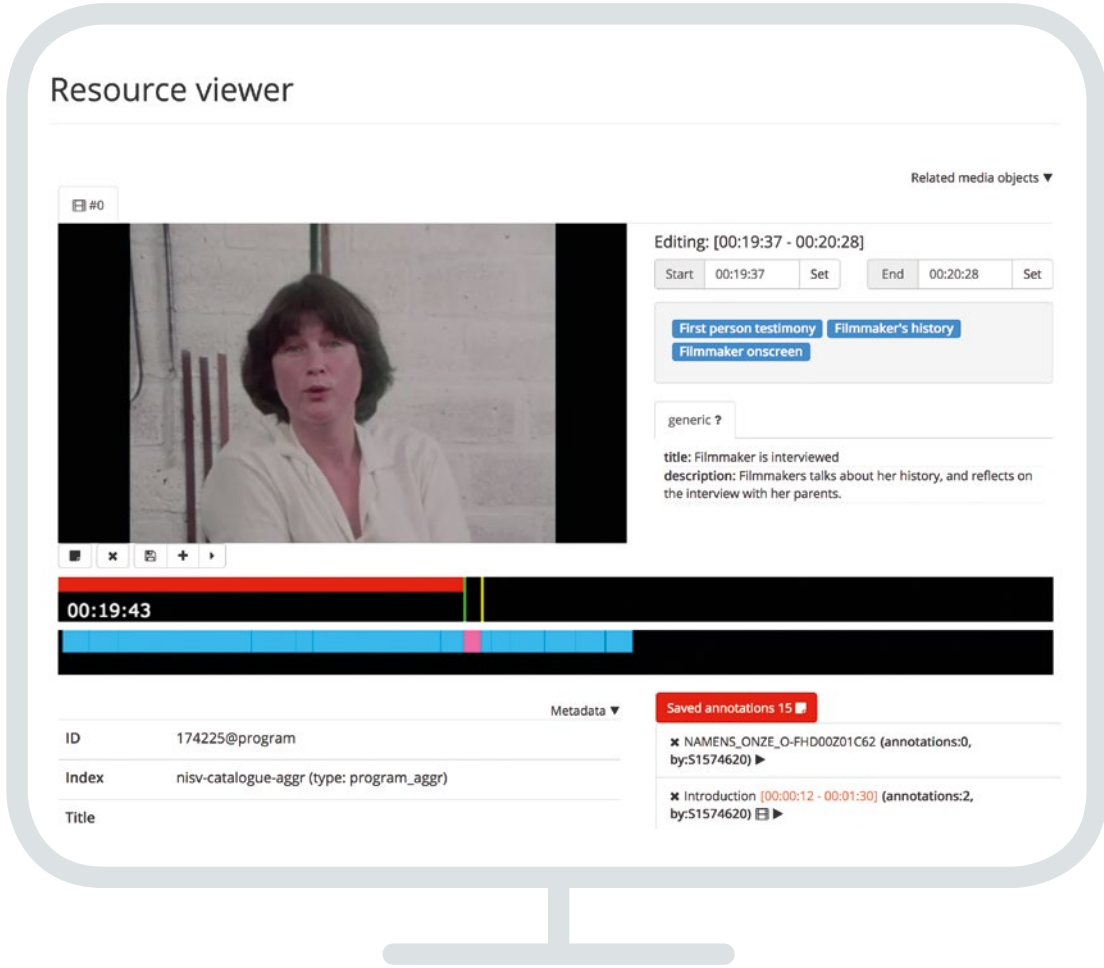
TRACING FIRST PERSON IN DOCUMENTARY HISTORY IN AV-COLLECTIONS

UNIVERSITY OF GRONINGEN,
UNIVERSITY OF AMSTERDAM,
NETHERLANDS INSTITUTE OF SOUND AND VISION

CONTACT:
SUSAN AASMAN
S.I.AASMAN@RUG.NL

How to reconstruct the emergence of a particular genre in a large dataset of audiovisual material?

In order to trace a transformation from the traditional objective documentary as fair and fact minded towards one with an appreciation for a more personal and subjective style, the 'M&M'-team aimed to explore in the archives of the Netherland Institute for Sound and Vision a large corpus of Dutch documentaries that were produced for public broadcast in the period of 1960-1990. The research team experimented with and tested the search functionalities and more important, the suitability of a web-based video annotation tool for media historical research within the CLARIAH infrastructure and the MediaSuite in particular. The video annotation tool enables a user to segment a digital moving image file and add annotations (metadata) to these time-coded segments. The main discovery of this research was not so much about the historical transformation but more about the appropriate methodology: using video annotation needs a specific choice about what precisely can be a unit of analysis in a complex genre like first person documentary. This experiment learned us how to improve our research strategy, and in addition, it helped us to understand how to use and improve the video annotation tool.



FOR MORE: SEE THE VIDEO ANNOTATION TOOL SCREENCAST:
WWW.YOUTUBE.COM/WATCH?V=KL-YXK856OQ

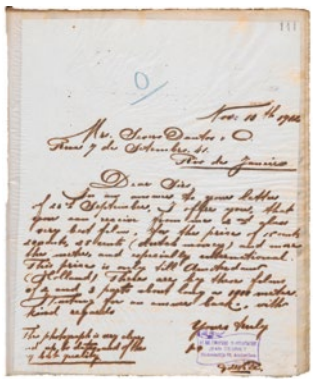
MIMEHIST

ANNOTATING EYE FILMMUSEUM'S JEAN DESMET COLLECTION

TOWARDS MIXED MEDIA ANALYSIS IN DIGITAL MEDIA HISTORY IN AV-COLLECTIONS

UNIVERSITY OF AMSTERDAM, NETHERLANDS
INSTITUTE OF SOUND AND VISION

CONTACT:
CHRISTIAN OLESEN
C.G.OLESEN@UVA.NL



CARBON COPY OF HANDWRITTEN LETTER BY JEAN DESMET'S ASSISTANT GEORGE DE VRÉE DATED NOV. 10TH 1912

The research project MIMEHIST: Annotating Eye's Jean Desmet Collection aimed at unlocking Eye Filmmuseum's digitized Jean Desmet Collection and facilitating scholarship on it with video annotation tools in the CLARIAH Media Suite. The Desmet Collection is a unique resource for media historians. It contains a large amount of rare films from silent cinema's transitional years and an extensive documentation of cinema exhibition and distribution practices from the early to mid-twentieth century. For these reasons the Collection is internationally renowned and also part of UNESCO's Memory of the World Register.

MIMEHIST has embedded the Collection's approximately 950 films produced between 1907 and 1916, 1050 posters and business archive containing around 127,000 documents from eight decades, in the Media Suite. To unlock the collection, MIMEHIST has performed handwriting recognition, OCR and made a visual classification of Desmet's business archive. This has improved the archive's searchability drastically and allows scholars to browse and annotate items - all in high resolution - with great ease. Consequently, the Desmet Collection is now much more accessible and can stimulate research on film distribution, exhibition and content in cinema's early years to a greater degree than hitherto possible.

CREDITS: EYE FILMMUSEUM

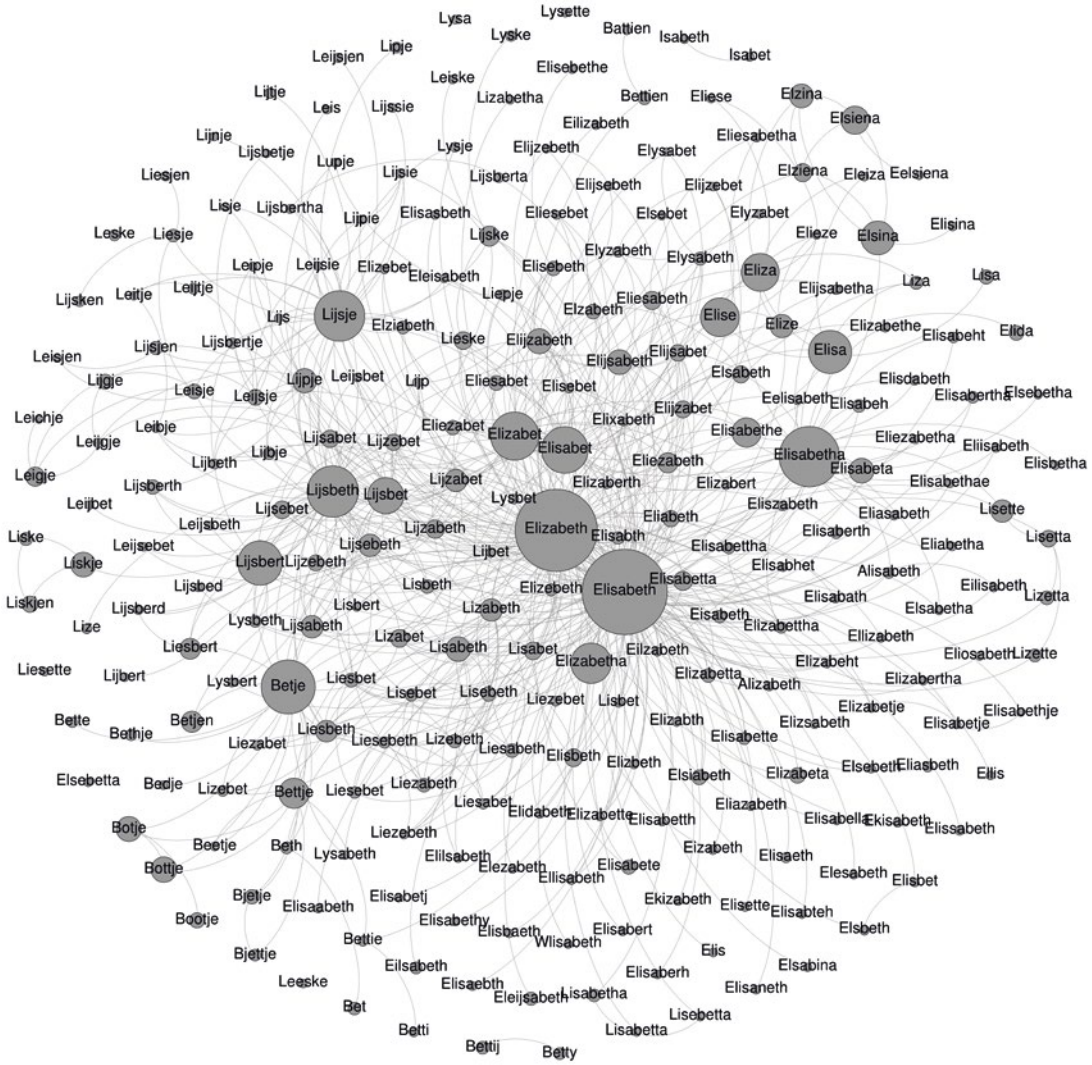
NAMES

DUTCH CORPUS OF PERSON NAME VARIANTS

UTRECHT UNIVERSITY, TILBURG UNIVERSITY,
DUTCH LANGUAGE INSTITUTE,
HUYGENS ING

CONTACT:
GERRIT BLOOTHOOFT
G.BLOOTHOOFT@UU.NL

Spelling variation, variants and digitization errors in person names are serious obstacles for search operations in historical documents. A solution could be the spelling standardization of surnames and given names. But ambiguities and alternative interpretations make this a non-trivial task which requires expert evaluation assisted by automatic analyses. The NAMES project aimed to standardize 564,000 different surnames and 190,113 different given names from 19th century sources with 52.5 million tokens with the help of the Clariah tool TICCL. A subset of these names was already automatically related to a standard as they could be identified as having been used for the same individual. This subset has been reviewed by experts which resulted in 127,154 surnames associated to 11,278 standards and 49,804 given names associated to 782 gender independent standards. Unfortunately, TICCL did not succeed to support the extension of this set. Instead, brute force comparison of the remaining names to names with a standard, and extending the number of standards, increased the coverage of standardized tokens to 99,43% for given names and 98,51% for surnames. Data will be made available in RDF format for linked open data and as Lexicon service. In addition, digital versions of name dictionaries will be made accessible.



VARIANT CLOUD OF ELISABETH WHERE EDGES DENOTE PROVEN VARIANT PAIRS.
THE SIZE OF A NODE IS PROPORTIONAL TO NAME FREQUENCY, WHICH IS >9 FOR THIS SET.

NarDis

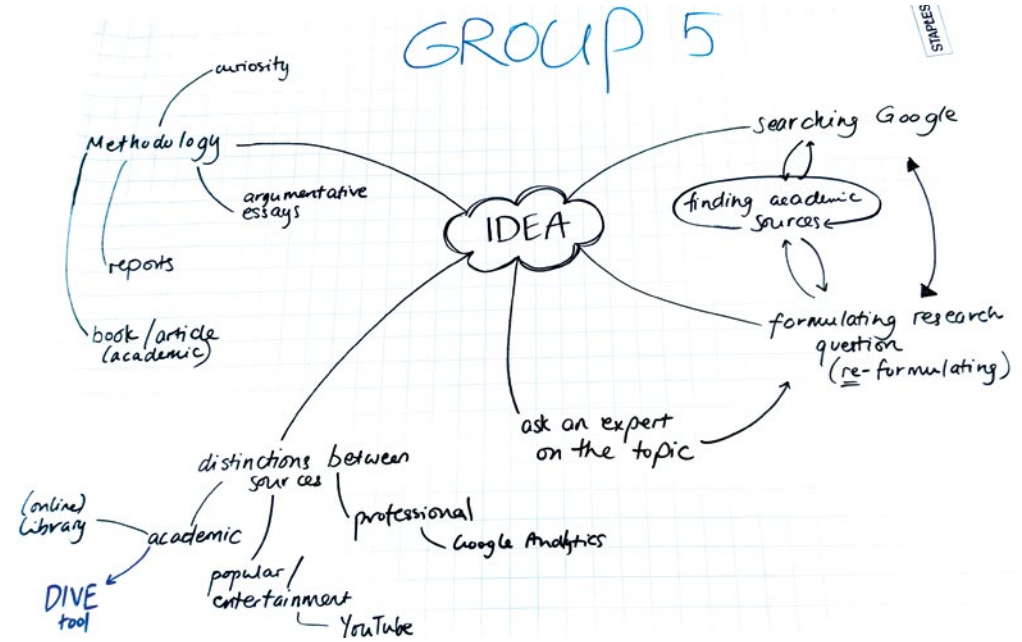
NARRATIVIZING DISRUPTION HOW EXPLORATORY

SEARCH CAN SUPPORT MEDIA RESEARCHERS TO INTERPRET ‘DISRUPTIVE’ MEDIA EVENTS AS LUCID NARRATIVES

UNIVERSITY OF AMSTERDAM,
UNIVERSITY OF GRONINGEN, VU UNIVERSITY,
NETHERLANDS INSTITUTE OF SOUND AND VISION

CONTACT:
SABRINA SAUER, S.C.SAUER@RUG.NL;
BERBER HAGEDOORN, B.HAGEDOORN@RUG.NL

This project investigates how CLARIAH’s exploratory search and linked open data (LOD) browser DIVE+ supports media researchers to construct narratives about events, especially ‘disruptive’ events such as terrorist attacks and natural disasters. This project approaches this question by conducting user studies to examine how researchers use and create narratives with exploratory search tools, particularly DIVE+, to understand media events. These user studies were organized as workshops (using co-creation as an iterative approach to map search practices and storytelling data, including: focus groups & interviews; tasks & talk aloud protocols; surveys/questionnaires; and research diaries) and included more than 100 (digital) humanities researchers across Europe. Insights from these workshops show that exploratory search does facilitate the development of new research questions around disruptive events. DIVE+ triggers academic curiosity, by suggesting alternative connections between entities. Beside learning about research practices of (digital) humanities researchers and how these can be supported with digital tools, the pilot also culminated in improvements to the DIVE+ browser. The pilot helped optimize the browser’s functionalities, making it possible for users to annotate paths of search narratives, and save these in CLARIAH’s overarching, personalised, user space. The pilot was widely promoted at (inter)national conferences, and DIVE+ won the international LODLAM (Linked Open Data in Libraries, Archives and Museums) Challenge Grand Prize in Venice (2017).



EXPLORATORY SEARCH VISUALIZED AS A MINDMAP IN USER STUDIES SESSIONS, USING CO-CREATION AS A METHOD FOR MAPPING SEARCH PRACTICES AND STORYTELLING DATA (DISCUSSED FURTHER IN HAGEDOORN & SAUER, FORTHCOMING 2018/2019)

OpenGazAm

LINKED OPEN DATA GAZETTEERS OF THE AMERICAS

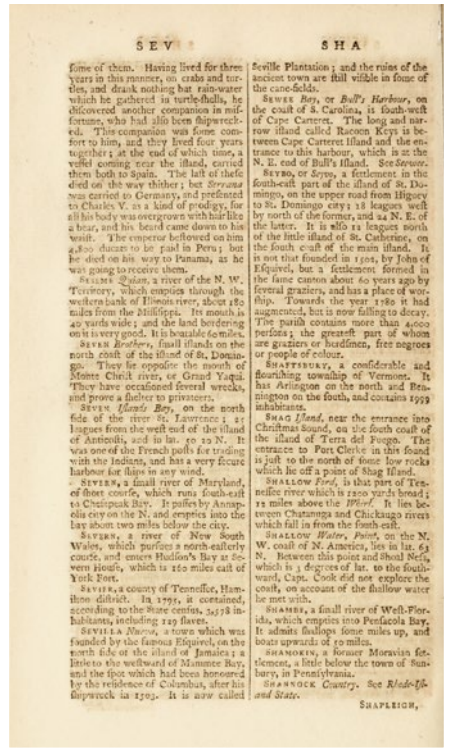
INTERNATIONAL INSTITUTE FOR SOCIAL HISTORY,
TILBURG UNIVERSITY, HUYGENS ING,
VU AMSTERDAM

CONTACT:
ROMBERT STAPEL
ROMBERT.STAPEL@IISG.NL

In 1797, Boston-based geographer Jedidiah Morse published the first edition of his momentous 'The American Gazetteer'.
It includes around 7,000 unique place name descriptions in the newly founded United States and in the European colonies in both North and South America, and the Caribbean.

The American Gazetteer provides a unique contemporary view of the Early Modern American contents. Its entries range from just a couple of words to several pages and contain basic information on the geographic location and administrative hierarchies of the localities, as well as descriptive notes. Much emphasis is placed on distances, navigability of waterways, types of traded commodities, climates, facilities, and so forth – all relevant for merchants seeking new fortunes.

The goal of the OpenGazAm-project is to create a Linked Open Data Gazetteer that will be interoperable with the *World Historical Gazetteer* and *Pelagios*. Digital historical gazetteers such as The American Gazetteer are indispensable in modern humanities research. Existing non-historical digital gazetteers, such as GeoNames, have much difficulty in identifying and disambiguating historical toponyms. Spelling variations, changing place names, and discontinued localities, are omnipresent in historical sources and hamper quick and easy identification of places.



SEV SHA

some of them. Having lived for three years in this manner, on crabs and turtles, and drank nothing but rain-water which he gathered in turtle-shells, he discovered another companion in misfortune, who had also been shipwrecked. This companion was some comfort to him, and they lived four years together; at the end of which time, a vessel coming near the island, carried them both to Spain. The last of these died on the way thither; but *Stevens* was carried to Germany, and presented to Charles V. as a kind of prodigy, for all his body was overgrown with hair like a bear, and his beard came down to his waist. The emperor bestowed on him 4,800 ducats to be paid in Peru; but he died on his way to Panama, as he was going to receive them.

SEVERN *Queen*, a river of the N. W. Territory, which empties through the western bank of Illinois river, about 180 miles from the Mississipi. Its mouth is 40 yards wide; and the land bordering on it is very good. It is boatable 60 miles.

SEVEN *Brothers*, small islands on the north coast of the island of St. Domingo. They lie opposite the mouth of Monte Christ river, or Grand Yaqui. They have occasioned several wrecks, and prove a shelter to privateers.

SEVEN *Islands Bay*, on the north side of the river St. Lawrence; 21 leagues from the west end of the island of Anticosti, and in lat. 50 19 N. It was one of the French ports for trading with the Indians, and has a very secure harbour for ships in any wind.

SEVERN, a small river of Maryland, of short course, which runs south-east to Chesipeak Bay. It passes by Annapolis city on the N. and empties into the bay about two miles below the city.

SEVERN, a river of New South Wales, which purfues a north-eaſterly courſe, and enters Hudſon's Bay at Severn Houſe, which is 160 miles eaſt of York Fort.

SEVIER, a county of Tenneſſee, Hamilton diſtrict. In 1795, it contained, according to the State cenſus, 3,578 inhabitants, including 129 ſlaves.

SEVILLA *Nueva*, a town which was founded by the famous Eſquivel, on the north ſide of the iſland of Jamaica; a little to the weſtward of Manmee Bay, and the ſpot which had been honoured by the reſidence of Columbus, after his ſhipwreck in 1503. It is now called Seville Plantation; and the ruins of the ancient town are ſtill viſible in ſome of the cane-fields.

SEWEE *Bay*, or *Bull's Harbour*, on the coaſt of S. Carolina, is ſouth-weſt of Cape Carteret. The long and narrow iſland called Racoon Keys is between Cape Carteret Iſland and the entrance to this harbour, which is at the N. E. end of Bull's Iſland. See *Seewee*.

SEYBO, or *Seyne*, a ſettlement in the ſouth-eaſt part of the iſland of St. Domingo, on the upper road from Higuey to St. Domingo city; 18 leagues weſt by north of the former, and 24 N. E. of the latter. It is alſo 12 leagues north of the little iſland of St. Catherine, on the ſouth coaſt of the main iſland. It is not that founded in 1502, by John of Eſquivel, but a ſettlement formed in the ſame canton about 60 years ago by ſeveral graziers, and has a place of worſhip. Towards the year 1780 it had augmented, but is now falling to decay. The parith contains more than 4,000 perſons; the greateſt part of whom are graziers or huſbandmen, free negroes or people of colour.

SHAFTSBURY, a conſiderable and flouriſhing townſhip of Vermont. It has Arlington on the north and Benningtton on the ſouth, and contains 1959 inhabitants.

SHAG *Iſland*, near the entrance into Chriſtmas Sound, on the ſouth coaſt of the iſland of Terra del Fuego. The entrance to Port Clerke in this ſound is juſt to the north of ſome low rocks which lie off a point of Shag Iſland.

SHALLOW *Ford*, is that part of Teneſſee river which is 1200 yards broad; 12 miles above the *Whirl*. It lies between Chatanuga and Chickaugo rivers which fall in from the ſouth eaſt.

SHALLOW *Water*, *Point*, on the N. W. coaſt of N. America, lies in lat. 63 N. Between this point and Shoal Neſs, which is 3 degrees of lat. to the ſouthward, Capt. Cook did not explore the coaſt, on account of the ſhallow water he met with.

SHAMER, a ſmall river of Weſt-Florida, which empties into Penſacola Bay. It admits ſhallops ſome miles up, and boats upwards of 50 miles.

SHAMOKIN, a former Moravian ſettlement, a little below the town of Sunbury, in Penſylvania.

SHANNOCK *Country*. See *Rhode-Iſland State*.

SHAPLEIGH,

PAGE FROM THE AMERICAN GAZETTEER (1797).

In order to proceed from image to text and from text to structured data, the project uses the Handwritten Text Recognition toolkit *Transkribus*, further enhanced by applying TICCL, which has been specially tested and adapted for this project. Linked Data conversion tools developed in CLARIAH WP4 are used for converting our geodata into Linked Open Data.

ReSpoNs

REMEDIATION IN SPORTS NEWS

UNIVERSITY OF GRONINGEN, UTRECHT UNIVERSITY,
NETHERLANDS INSTITUTE FOR SOUND AND VISION,
NATIONAL LIBRARY OF THE NETHERLANDS

CONTACT:
MARCEL BROERSMA
M.J.BROERSMA@RUG.NL

In media history it is often assumed that the rise of television as a (journalistic) medium has had a considerable influence on how newspapers covered the news. The popularity of television coverage which offered liveness and a visual experience, forced newspaper journalism to rethink their ways of reporting. Yet, remediation of these media has never been studied empirically.

RESPONS AIMS TO:
a) analyze processes of remediation between newspapers and television between 1959 and 1989;
b) test and further develop the functionalities of the comparative search tool.

The first step entailed research for a demonstration scenario based on end-user experiences with the comparative search tool. It outlines how the tool should ideally look like, determining its prerequisite features. This resulted in a ‘wish list’ with features for the media suite. The demonstration scenario has been continuously updated during the project to add new insights. Unfortunately, the digital newspaper data were only added to the media suite in the last month of the project. Despite this setback, we analyzed remediation between newspapers and television by directly using the digital collections of the KB and ISV. This resulted in a paper on the newspaper discourse on televised sports, and we are currently working on a paper on the way newspaper coverage developed under influence of the rise of television.

SERPENS

SEARCH PEST AND NUISANCE SPECIES

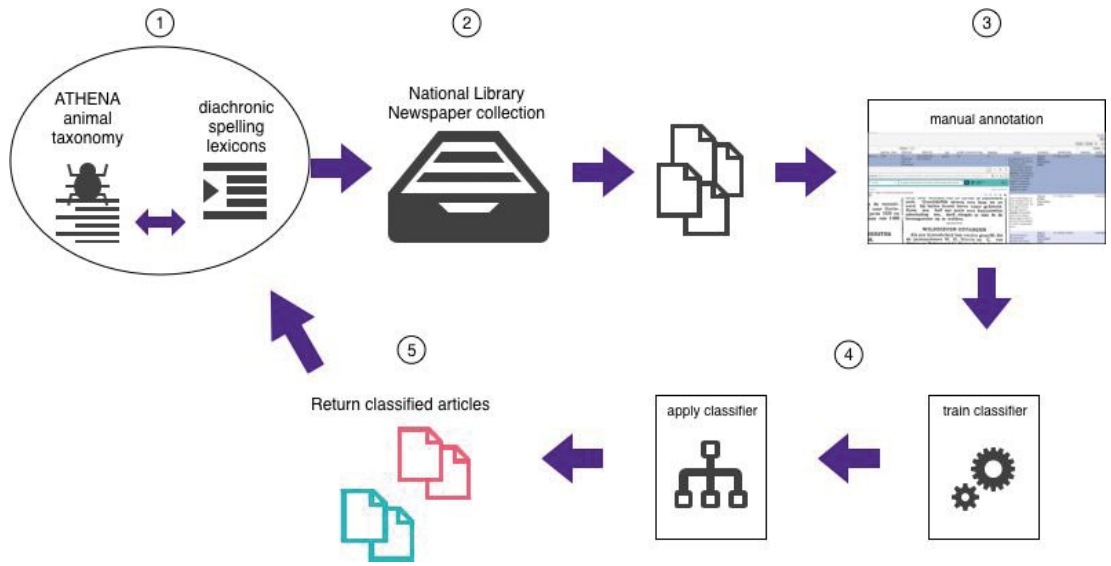
CONTEXTUAL SEARCH AND ANALYSIS OF PEST AND NUISANCE SPECIES THROUGH TIME IN THE KB NEWSPAPER COLLECTION

RADBOUD UNIVERSITY NIJMEGEN, KNAW HUMANITIES CLUSTER,
DUTCH LANGUAGE INSTITUTE,
NATIONAL LIBRARY OF THE NETHERLANDS, HUYGENS ING

CONTACT:
ROB LENDERS
R.LENDERS@SCIENCE.RU.NL

Historical newspapers are a fascinating source of information for historical ecologists to study interactions between humans and animals through time and space. Digitized newspaper archives are particularly interesting to analyze because of their breadth and depth and easy access. However, the size and the occasional noisiness of such archives also brings difficulties, as manual analysis still remains cumbersome and laborious. In SERPENS, we performed experiments to automate query expansion and categorization for the perception of alleged pest and nuisance animal species mentioned in digitized newspapers from a subset of the KB newspaper collection (1800-1940). We particularly focused on the perception of Mustelid species like polecats, martens and stoats. For animal taxonomy we made use of ATHENA; for query expansion we used lexicons; for categorization of newspaper articles we trained a Support Vector Machine model. Our results indicate that – with a rather limited number of training examples – we can fairly easily distinguish newspaper articles that are about animal species from those that are not (~92% accuracy) and between different types of subcategories of newspaper articles (e.g., articles about material damage caused by pest species, non-material damage, pest control and hunting; ~84% accuracy). Automated procedures like this can greatly enhance the usability of large digitized collections, not only for historical ecology but also for other fields in the natural sciences and humanities.

POLECAT. SOURCE: RIJKSSTUDIO



WORKFLOW SERPENS

ADAH PROJECTS

The eScience Center and CLARIAH have initiated four projects that will pursue new scientific domain challenges and enhance and accelerate the process of scientific discovery within the arts and humanities using computer science, data science, and eScience technologies.

The projects are collaborations with research teams from multiple Dutch academic groups.
The granted projects will use, adapt, and integrate existing methods and tools, as made available through the CLARIAH and eScience Center software infrastructures. Newly developed tools will be made available through the eScience Technology Platform of the Netherlands eScience Center and the CLARIAH Infrastructure for potential use in other studies. These projects will finish in the course of 2019.

BRIDGING THE GAP

DIGITAL HUMANITIES AND THE ARABIC-ISLAMIC CORPUS

PI: Christian Lange and Melle Lyklema (Utrecht University)

Despite some pioneering efforts in recent times, the computational analysis of Islamic intellectual history remains a largely unexplored field of research. Researchers still tend to study a narrow canon of texts, made available by previous Western researchers of the Islamic world largely based on considerations of the relevance of these texts for Western theories, concepts and ideas. Indigenous conceptual developments and innovations are therefore insufficiently understood, particularly as concerns the transition from premodern to modern thought in Islam.

This project harnesses state-of-the art Digital Humanities approaches and technologies to make pioneering forays into the vast corpus of digitised Arabic texts (ca. 10 times the size of the 'classical' Greek and Latin corpus) that has become available in the last decade. This is done along the lines of primarily two case studies, each of which examines a separate genre of Arabic and Islamic literary history: Islamic jurisprudence; and the Arabic literature on proselytism. By way of 'distant reading', these two corpora are studied in terms of the semantic shifts they gradually underwent (from the 8th to the 20th c.), and the terminological and conceptual differences obtaining between different clusters of texts within the corpus (e.g. the different schools of law in Islam, that is, the four major Sunni schools and the Shi'i school).

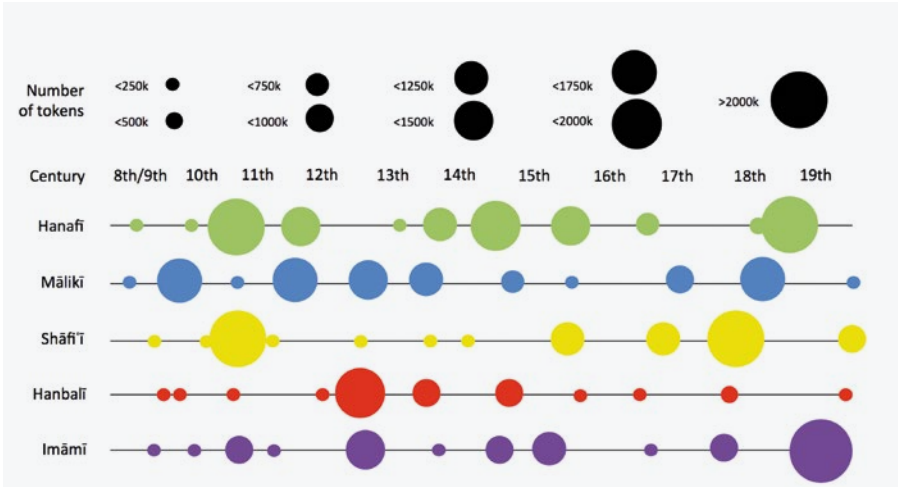


FIGURE: OVERVIEW OF THE UTRECHT BASED DIGITAL CORPUS OF ISLAMIC JURISPREDENCE

This project has developed an openly accessible, Arabic-compatible version of the corpus search engine BlackLab (based on Apache Lucene) that enables easy access to the two marked-up corpora and offers a set of tools for Arabic text mining and computational analysis. The project is inserted into an ongoing ERC project on Islamic intellectual history housed at the Department of Philosophy and Religious Studies at Utrecht University, and has collaborated closely with international initiatives in the field of Arabic Digital Humanities, culminating in the organisation of a KNAW academy colloquium, 'Whither Islamicate Digital Humanities? Analystics, Tools, Corpora' (13-15 December 2018).

EviDENce

EGO DOCUMENTS EVENTS MODELLING

HOW INDIVIDUALS RECALL MASS VIOLENCE

PI: Susan Hogervorst (Open University)

Much of our historical knowledge is based on oral or written accounts of eyewitnesses, particularly in cases of war and violence, when regular ways of documentation and record keeping are often absent. EviDENce studies how eyewitnesses have reported on violence, and how this may have changed over time. We use a collection of nearly 500 oral history interview transcripts about the Second World War (Getuigen Verhalen, stored at DANS) as well as the ego-documents (diaries, memoires, letters, autobiographies) available in Nederlab, covering a time span of 5 centuries.

Whereas humanities scholars are good at assessing texts for their relevance in relation to a particular topic or research question such as this, automating this assessment process, for example for distant reading or creating large corpora, is known to be problematic, especially when it comes to implicit mentions. EviDENce compares existing NLP methods to detect fragments containing mentions of such an ambiguous concept as violence, in a way that meets the standards of historical research.

TICCLAT

TEXT INDUCED CORPUS CORRECTION AND LEXICAL ASSESSMENT TOOL

PI: Martin Reynaert (Tilburg University)

The Text-Induced Corpus Clean-up tool TICCL, integral part of the CLARIN infrastructure, is globally unique in utilizing the corpus-derived word form statistics to attempt to fully-automatically post-correct texts digitized by means of Optical Character Recognition.

The NWO ‘Groot’ project Nederlab has delivered a uniformly processed and linguistically enriched diachronic corpus of Dutch containing an estimated 5-6 billion word tokens. We aim to extend TICCL’s correction capabilities with classification facilities based on specific data collected from the full Nederlab corpus: word statistics, document and time references and linguistic annotations, i.e. Part-of-Speech and Named-Entity labels. These data will complement a solid, renewed basis composed of the available validated lexicons and name lists for Dutch.

In this, TICCL as a post-correction tool will be transformed into TICCLAT, a lexical assessment tool capable of delivering not only correction candidates, but also e.g. more accurately dated diachronic Dutch word forms, more securely classified person and place names. To achieve this on scale, the TICCLAT project relies on a successful extension of TICCL’s anagram hashing towards text-induced morphological classification. TICCLAT’s capabilities will also be evaluated in comparison to human performance by an expert psycholinguist.

The data collected will be exportable for storage in a data repository, as RDF triples, for broad reuse. The project will greatly contribute to a more comprehensive overview of the lexicon of Dutch since its earliest days and of the person and place names that share its history. Its partners are the Dutch experts in lexicology, person names and toponyms.

NEWSGAC

NEWS GENRES

ADVANCING MEDIA HISTORY BY TRANSPARANT AUTOMATIC GENRE CLASSIFICATION

PI: Marcel Broersma (University of Groningen)

This project studies how genres in newspapers and television news can be detected automatically using machine learning in a transparent manner. This enables us to capture the often hypothesized but, due to the highly time-consuming nature of manual content analysis, largely understudied shift from opinion-based to fact-centred reporting. Moreover, we open the black box of machine learning by comparing, predicting and visualizing the effects of applying various algorithms on heterogeneous data with varying quality and genre features that shift over time. This enables scholars to do large-scale analyses of (historic) texts and other media types as well as critically evaluate the methodological effects of various machine learning approaches.

This project brings together expertise of journalism history scholars (University of Groningen), specialists in data modelling, integration and analysis (CWI), digital collection experts (National Library & Netherlands Institute for Sound and Vision) and e-science engineers (eScience Centre). It uses a big manually annotated dataset (VIDI-project PI) to develop a transparent and reproducible approach to train an automatic classifier. Building upon this, the project generates three outcomes:

1. A study that revises our current understanding of the interrelated development of genre conventions in print and television journalism based upon large-scale automated content analysis via machine learning;

2. Metrics and guidelines for evaluating the bias and error of the different pre-processing and machine learning approaches and of-the-shelf software packages;

3. A dashboard that integrates, compares and visualises different algorithms and underlying machine learning approaches which can be integrated in the CLARIAH Media Suite.

INTEGRATION PROJECTS

ATM

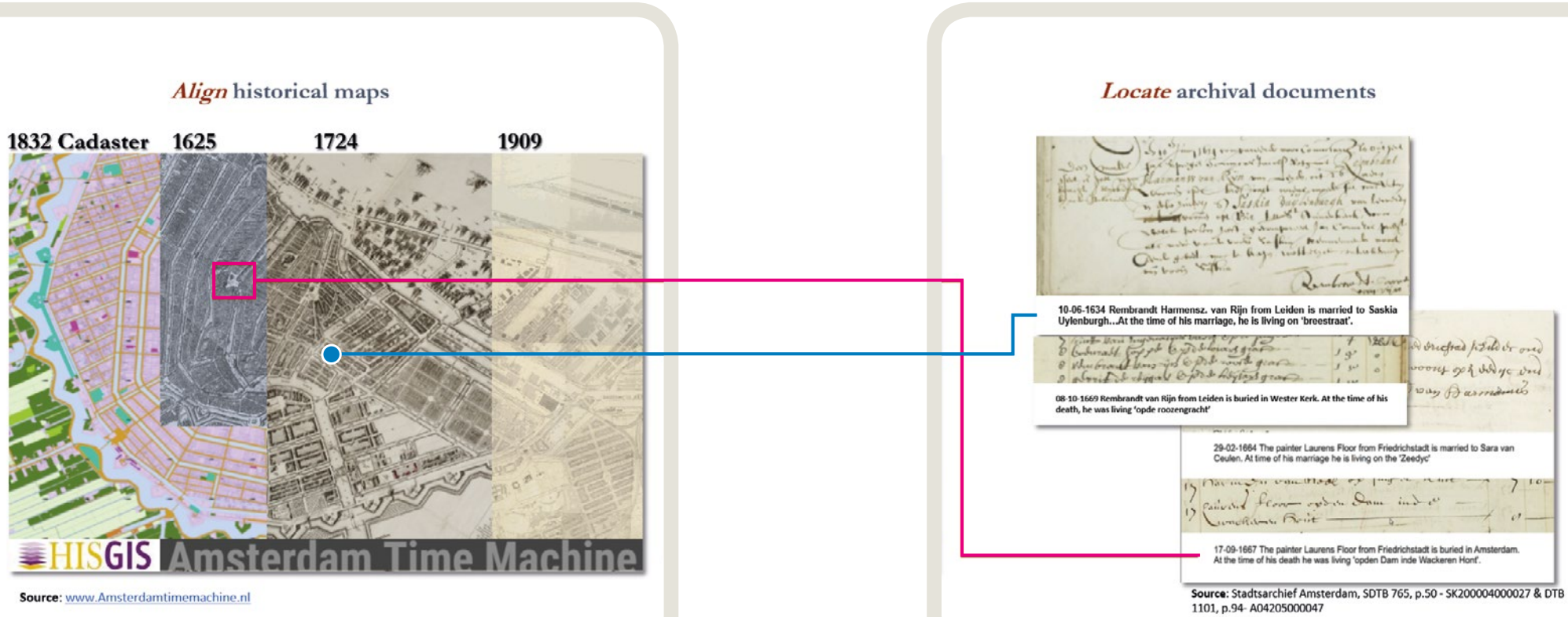
AMSTERDAM TIME MACHINE

SPATIAL HUMANITIES IN THE CLARIAH INFRASTRUCTURE

UNIVERSITY OF AMSTERDAM, FRYSKE AKADEMY,
KNAW HUMANITIES CLUSTER,
INTERNATIONAL INSTITUTE FOR SOCIAL
HISTORY, MEERTENS INSTITUTE, ADAMNET

CONTACT:
JULIA NOORDEGRAAF,
J.J.NOORDEGRAAF@UVA.NL
AMSTERDAMTIMEMACHINE.NL

Is it possible to travel back in time and walk the streets of historical Amsterdam? We certainly think so. The Amsterdam Time Machine (ATM) is an integrated platform to present historical information about people, places, relations, events, and objects in its spatial and temporal context. The web of data on the history of Amsterdam is created by systematically linking existing datasets from social and humanities research with municipal and cultural heritage data. Where possible this is done in the form of Linked Open Data. The linked data can then be organized and presented in spatial representations, such as geographical and 3D visualizations. The result is a 'Google Earth' for the past, which invites users to explore the city through space and time, at the level of neighborhoods, streets, or individual houses. ▶

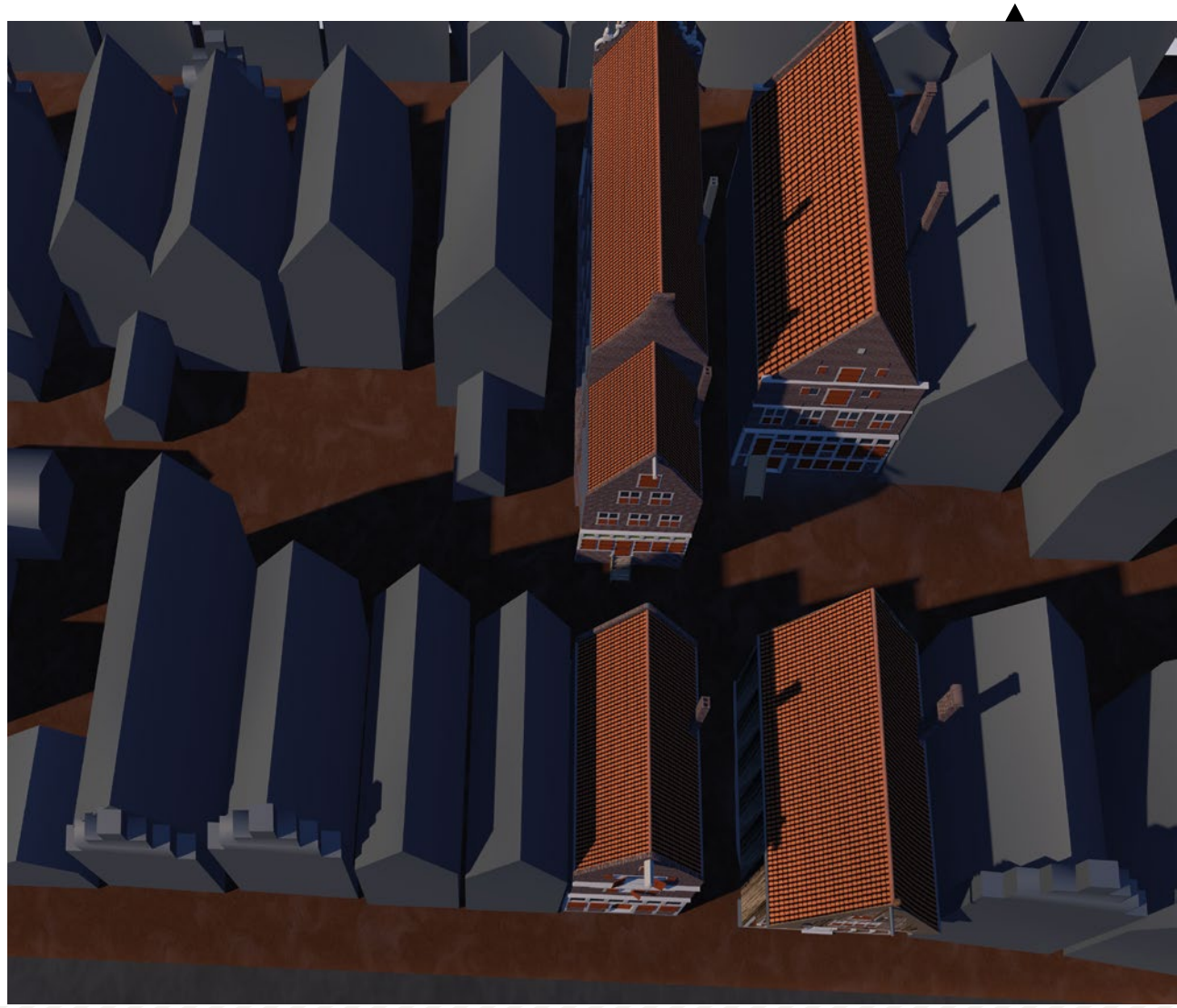


DEEP MAPPING CREATIVE AMSTERDAM OVER TIME (BY WEIXUAN LI).

A CLARIAH grant made it possible to develop a first proof of concept. In the CLARIAH Amsterdam Time Machine project the linked data from cultural heritage institutions made available in the AdamLink project is combined with that of various scholarly research projects at the International Institute for Social History, Huygens ING, Meertens Institute and University of Amsterdam, and integrated with a GIS developed by Fryske Akademy. Subsequently, the historical geographical and topological context for these linked datasets is made available open access in the CLARIAH infrastructure at the KNAW Humanities Cluster. The project also comprises three research use cases on language, social mobility and leisure. These use cases demonstrate how the Amsterdam Time Machine offers instruments for research into urban space as a connecting factor for observing and analyzing social and cultural processes. On the one hand, they testify to the potential of the framework for innovating disciplinary research in Linguistics, History and Media Studies. On the other hand, they show how the research infrastructure also supports interdisciplinary research, by making a connection between the social development of Amsterdam's historical population groups, their language development and their leisure activities in local theatres and cinemas.

More generally, ATM facilitates 'scalable digital humanities research': smoothly navigating historical data from the micro level of one location, anecdote or document to the macro level of patterns in large, linked datasets that expose broader social and cultural processes. Charles Tilly described the city as a "privileged site for study of the interaction between large social processes and routines of local life" (Tilly 1996, 704). The Time Machine operationalizes this by investigating the urban history of Amsterdam on a scale that varies between the micro level of a plot, person or place and the macro level of broader societal processes in the city as a whole - a microscope and telescope in one. Such a research environment offers an unprecedented opportunity to explore the relationship between physical and social space and how this connection was experienced and transformed over time. With space as a connecting factor, the Time Machine provides a concrete illustration of the research potential of linking social and economic data with cultural data, allowing researchers to study specific historical and cultural phenomena against the background of broader societal developments.

3D MODEL OF A MERCHANT'S HOME ('T PARADIJS) IN THE KALVERSTRAAT. AMSTERDAM, EARLY 16TH C (BY MADELON SIMONS, LOES OPGENHAFFEN ET AL.).

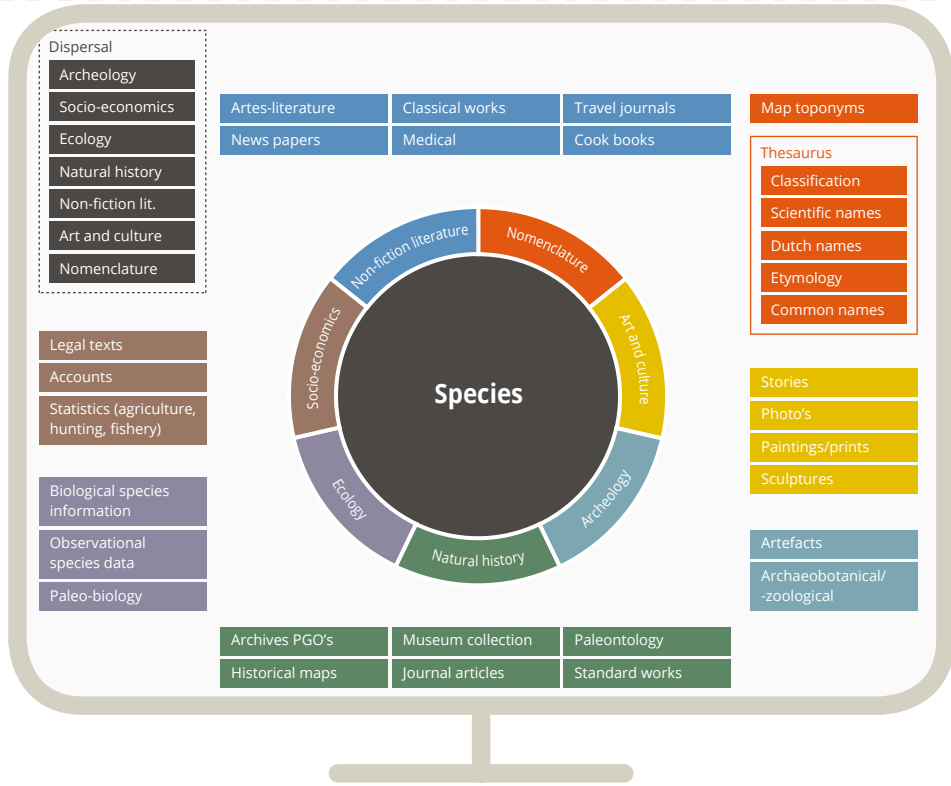


ATHENA

ACCESS TOOL FOR DATA ON HISTORICAL ECOLOGY AND ENVIRONMENTAL ARCHEOLOGY

RADBOUD UNIVERSITY NIJMEGEN

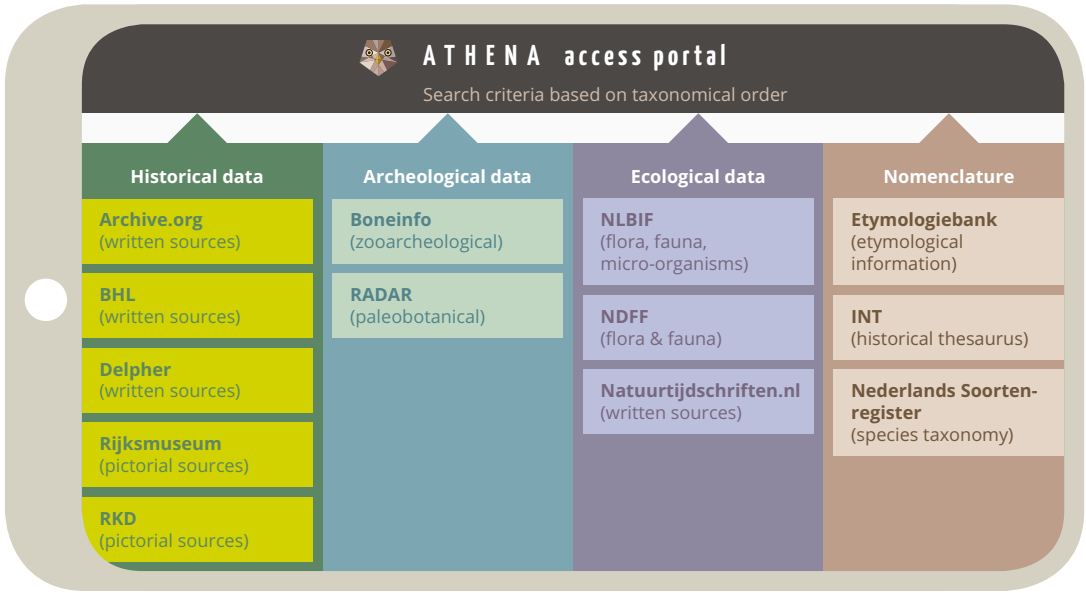
CONTACT: THOMAS GOETHEM,
TGOETHEM@SCIENCE.RU.NL
WWW.ATHENA-RESEARCH.ORG



How did humans shape the environment and vice versa? Why did some plants and animals disappear from the Netherlands? With growing concerns about the future of nature and the sustainable development of human societies, it is very important to study these issues in a historical perspective and from all angles. To that end, the interdisciplinary ATHENA project has been developed.

ATHENA is a digital research platform that brings together a wide range of historical, bio-archaeological and biological sources that relate to the historical occurrence of vascular plants and vertebrates in the Netherlands. Examining the historical relationship between man and nature requires almost by definition an interdisciplinary approach. For example, a known problem in studying the (historical) influence of humans on the environment is the lack of quantitative information from before 1950 about the occurrence of plants and animals. However, studying historical newspaper articles, floras, encyclopaedias, herb books, bestiaries, archaeological excavations of bone remains and botanical macro rests, but also relevant prints and paintings, can fill this gap. Currently, such data and datasets are scattered or even lacking. Moreover, the data that is available is generally missing historical, social, cultural and ecological context. To address this situation, historians, archaeologists and ecologists have come together to share and combine data in the innovative ATHENA research platform.

ATHENA builds on the model of data hubs, creating a network of databases that provide state-of-the-art data for their own disciplinary fields of research. The data in ATHENA can be classified in different data types: textual sources, media objects (images) and (semi-) structured data, following the three leading disciplines of CLARIAH nicely. The different datasets can be accessed in a unified manner through a single data portal. However, linking and aggregating heterogeneous, but contextually related, information to create a body of knowledge on historical human-nature relationships is no small feat.
To facilitate innovative forms of cross-database queries and information retrieval, the ATHENA project has made use of database technologies developed in CLARIAH. The datasets brought together in the Athena project have been imbedded in the common CLARIAH research infrastructure, by adopting the CLARIAH data formats for each of the data types.



This made it possible to expand the ATHENA network of databases with relevant data from the broader CLARIAH research network. For example, the natural history thesaurus for plants and animals has been expanded with dialect and spelling variations, synonyms and dating information, enabling much deeper queries of textual sources. Moreover, research technologies developed by the three leading CLARIAH disciplines have been applied. For example, a computational linguistics approach, in combination with the thesaurus, has been used to create a semantically enriched dataset of newspaper articles from the KB newspaper corpus. Another example is the enrichment of pictorial sources, such as paintings and prints, by making use of annotation and crowdsourcing tools developed specifically for media studies.

The ATHENA project shows, in part, the potential of the common research infrastructure developed in CLARIAH for the humanities and beyond. Applying state of the art tools has enabled to examine and (re)contextualize knowledge on historical human-nature relationships from heterogeneous digital data sources. The platform provides access to information that was often limited

available and not jointly investigated. Users are alerted to new and surprising sources, which in turn may lead to new research questions. ATHENA is an internationally unique research platform that will offer researchers from the humanities and natural sciences, but also policy makers, NGO's and the broader public, a tool for exploring major questions in the field of human-nature relationships.

DIGIFIL
DIGITAL FILM LISTINGS

Kaspar Beelen, Kathleen Lotze, Ivan Kisjes en Thunnis van Oort (University of Amsterdam)

DIGIFIL aims to digitise the Dutch Filmladders (the weekly listings of movie showtimes at local cinema theatres or other venues) and contextual information about the wider movie landscape as reported in historical newspapers (such as movie reviews and descriptions). The screenings constitute the focal point of film culture: they are the place where distributors, exhibitors and audiences meet. Collecting information about these encounters, and embedding them in their wider discursive context, yields an invaluable resource for linguists, socio-economic historians and media scholars to study the ways in which cinema-going contributed to the formation of modern societies.

DIGIFIL continues and extends the work of film historian Karel Dibbets on 'Cinema Context', an online database that emerged out of an NWO-funded project. Cinema Context comprises digitised programming data until 1940, as well as extensive information about distributors, cinema theatres and the people and companies behind them. The database stands out internationally as one of the earliest examples of open access and has greatly facilitated data-driven research in media history.

DIGIFIL builds upon the digitization effort of the National Library. Their current collection, available via Delpher, already contains an impressive set of digitized, segmented and enriched newspapers. The point of DIGIFIL is to improve digitization and enrichment of specific sections in the newspaper corpus. The participants use the available digitized materials as a starting point but refine and extend them wherever that is required, using existing tools developed by CLARIAH Work Package 3 (PICCL, TICCL, FROG). DIGIFIL demonstrates how semi-structured text

converted to entries in a database. It showcases the power of computational techniques for extending existing databases such as Cinema Context. In doing so, this project provides scholars who work on similar sources with various computational tools and best-practices to transform their own sources to machine readable and manipulable data.

CLARIAH TOOLS

User-friendly tools are an important part of the CLARIAH infrastructure. Tools with which scholars, students and other interested parties can research, edit, combine and save both their own and other data. A selection of the tools that are ready for use are described below. In the coming years, the CLARIAH-PLUS project will output more tools; an overview will be available on the CLARIAH website and the tools themselves will be available on http://github.com/CLARIAH.

Name	AutoSearch
Why	AutoSearch allows non-technical users to upload, search and analyze linguistically annotated corpus data.
Description	AutoSearch allows CLARIN-authenticated users to upload corpora to a private workspace, after which the corpus data can be searched and analyzed in a user-friendly web application. Corpora should (preferably) be annotated at the token level with properties such as part of speech, lemma and word form. Supported corpus file formats are, among others, FoLiA, TEI, CHAT, EAF, tab-separated files. New file formats can be defined by the user by editing a configuration file. The search application is powered by the INT BlackLab corpus search engine; the user interface is similar to the interface of the OpenSoNaR corpus application. For Dutch corpora without annotations, one can first use the PICCL workflow for OCR, post-correction and linguistic annotation to enrich plain text with annotations for lemma and part of speech in FoLiA format.
Application data	Text corpus files with linguistic annotation in various formats (TEI, FoLiA, CHAT, ...)
Application area	Corpus exploitation
Link to software	Web application: https://portal.clarin.inl.nl/autocorp Source: https://github.com/INL/corpus-frontend

Name	CLARIAH Authentication Service
Why	The CLARIAH authentication service removes the complexities of securing web services. Software engineers can focus on the development of research tools without having to worry about GDPR compliance, IdP administration etc.
Description	The CLARIAH authentication service offers both SAML and OpenIDC/OAUTH compliancy and filters the personal data transferred by IdP's on 'Legitimate Interest'. Data that is not needed by a Service Provider is deleted to comply with GDPR regulations. Moreover, the service identifies users across the CLARIAH infrastructure based on a persistent ID, not on an email address, as email addresses have a tendency to change over time. This will allow CLARIAH to also link different accounts to a single user, enabling people to login with multiple identities. Users who do not have access to an institutional account that is part of the CLARIN.EU managed federation of IDP's, can request a so-called 'homeless user' account at the CLARIN IDP and gain access to services through this channel. The CLARIAH authentication service is a stateless proxy running SATOSA. It is setup in such a manner to prevent it becoming a single-point of failure in the infrastructure.
Application data	-
Application area	Security, User Authentication, Authorization and Identification.
Link to software	http://authentication.clariah.nl

Name	CoW
Why	CSV files are very commonly used to store research data. However, very often researchers pose research questions that require to combine, clean, and harmonize multiple CSV files before conducting any analysis. Moreover, publishing these prepared CSVs on the Web to enable their reuse by other researchers is also troublesome, and often inefficient for large CSV files. Representing these CSVs files in the RDF data format as Linked Data would be a useful step to overcome these issues.
Description	CoW is a comprehensive and high-performance tool for batch conversion of multiple datasets expressed in CSV to RDF. In a first step, CoW creates a JSON schema file from the input CSV, expressed using an extended version of the W3C CSVW standard, which can be manually adjusted by the user to accommodate their needs: selecting specific columns, creating new virtual columns, combining the values of different columns to mint URIs, etc. In a second step, CoW uses the instructions in this JSON schema file to build an RDF file correspondingly.
Application data	CSV files, JSON schema files CSVW compliant
Application area	Creating Linked Data
Link to software	https://github.com/CLARIAH/COW https://github.com/CLARIAH/cattle
Comment	CoW is a command line application (CLI). CoW can be locally installed via pip with ‘pip install cow_csvw’. An online version of CoW, cattle, is available at http://cattle.datalegend.net Cattle can be locally installed via docker with ‘docker run clariah/cattle’. CoW stands on the shoulders of the QBer giant.

Name	Datastories
Why	Explaining the benefits of Linked Open Data (LOD) for researchers not familiar with these techniques can be difficult, and in the worst case the technicalities can discourage potential users. Datastories demonstrates how LOD can enhance (historical) research by showing live output of integrated SPARQL queries on historical datasets, presented within a narrative about the data presented. By clicking the arrow button above the SPARQL results, users can edit and expand queries to their own liking.
Description	Datastories provides live SPARQL query results of several historical linked open datasets within a story about the presented dataset(s).
Application data	Datastories queries a selection of linked open datasets (RDF), mostly from https://druid.datalegend.net/, but other endpoints are also supported. Queries (.rq) are stored at a Github repository.
Application area	Humanities, Linked Open Data presentation and querying
Link to software	http://datastories.datalegend.net https://github.com/CLARIAH/wp4-stories

Name	Druid
Why	To work with Linked Data we use various tools to create, browse and query Linked Data. Druid provides a place where these tools are connected and can be used in conjunction. Moreover, it provides a friendly way of storing and presenting datasets, including sparql endpoint generation. In addition to the obvious metadata, it provides example resources giving a better feel for the type and quality of the data.
Description	Druid allows one to create, browse, query, visualize and present Linked Data. It's like a working environment where you can start out with a CSV and end up with a 'connected' dataset, meaning that you're able to retrieve information from other datasets to augment your own with. Moreover, it allows you to present that Linked Data to others, technically via a SPARQL endpoint, or more verbally and visually, via data stories (see below) and Yasgui's mapping and visualization capabilities.
Application data	Data in nearly any type of Linked Data format (e.g. turtle, nquads) is accepted in Druid. Moreover, it provides a file storage, so any 'source' file can be stored as well. CSV's that are uploaded can be directly transformed into Linked Data using CoW.
Application area	Any field using Linked Data. The majority of the data in Druid are from the Humanities and Social Sciences.
Link to software	http://druid.datalegend.net
Comment	Druid is the product of an unique symbioses between a multi-disciplinary academic team and the company Triply

Name	FLAT
Why	The output of automated linguistic enrichment tools needs to be properly viewable for researchers. Moreover, researchers need tools to edit this output or linguistically annotate their own data manually, for instance in efforts to build a gold standard corpus or training data. FLAT allows researchers to do this by providing an environment to view, edit and create linguistically annotated documents. It is a tool both for individual annotators or larger annotator teams.
Description	FLAT is a web-based linguistic annotation tool that allows human annotators to view and edit linguistically annotated documents. FLAT builds on the FoLiA format which has been developed as part of CLARIN-NL and CLARIAH for over six years. A wide variety of linguistic annotation types is supported and the tool and underlying format are tagset and language independent. FLAT has been successfully employed in multiple annotation projects over the years.
Application data	FoLiA XML Documents
Application area	Linguistic Annotation
Link to software	https://github.com/proycon/FLAT

Name	FROG
Why	Automatic linguistic enrichment of Dutch text is necessary for a wide variety of applications.
Description	Frog is an integration of memory-based natural language processing (NLP) modules developed for Dutch. All NLP modules are based on Timbl, the Tilburg memory-based learning software package. Frog integrates a tokeniser called ucto [1], another CLARIAH project, and provides part-of-speech tagging, lemmatisation, morphological analysis, a dependency parser, a base phrase chunker, and a named-entity recognizer module. This make Frog a good solution for automatic linguistic enrichment of Dutch texts. Frog supports the FoLiA format, developed in CLARIAH, for both input and output.
Application data	Text in plain text or FoLiA XML format.
Application area	Enriching text corpora with linguistic annotations.
Link to software	https://languagemachines.github.io/frog [1] https://languagemachines.github.io/ucto

Name	GrETEL 4
Why	GrETEL enables researchers to search for grammatical constructions in terms of grammatical relations (subject, object, etc.) and syntactic categories (noun, verb, noun phrase, etc.). A key characteristics of GrETEL is that it enables a researcher to define such queries on the basis of an example sentence (Query By Example).
Description	GrETEL stands for Greedy Extraction of Trees for Empirical Linguistics. It is a user-friendly search engine for the exploitation of syntactically annotated corpora or treebanks, originally developed by KU Leuven. In CLARIAH-CORE a new version has been developed (Version 4). Key new features added in CLARIAH-CORE are: (1) one can upload one's own corpus, in a variety of formats; (2) an Analysis component has been added which enables analysis of the search results in combination with metadata.
Application data	Text corpora in plain text, TEI, FoLiA or CHAT format and treebanks and parsebanks in LASSY format.
Application area	Advanced enrichment, search and analysis of text corpora for linguistic research.
Link to software	http://gretel.hum.uu.nl/gretel4/ng/home

Name	grlc
Why	Knowledge Graphs and Linked Data are useful tools for integrating and publishing data on the Web, but using and accessing them is usually troublesome. It requires users to learn specific knowledge representation and query languages such as RDF and SPARQL. Linked Data APIs have been proposed as a method to ease this problem, at least for developers; however, building such APIs is often a repetitive and time consuming task, which often hides SPARQL queries under hard to maintain implementations.
Description	grlc is a middleware server that builds APIs on top of Linked Data automatically, alleviating both the developers' costs of manually writing these APIs, and the issues end users face when they want to access and query Linked Data sources. In order to do so, it leverages publicly shared and annotated SPARQL queries from the user community, mainly published in the software version control portal GitHub. grlc accesses these queries and builds APIs by reusing them automatically, and enables an easier and universal access to Linked Data using standard Web mechanisms.
Application data	SPARQL queries, TPF queries, JSON prototypes
Application area	Accessing Linked Data, APIs
Link to software	http://grlc.io/
Comment	Can be run on the Web without local installations at http://grlc.io. Can be run locally from Docker with 'docker run clariah/grlc'.

Name	Kaldi_NL Speech Recognition
Why	Metadata on AV is sparse. Speech recognition transcripts enable searching.
Description	KALDI is a well-known, 'state-of-the-art' speech recognition toolkit. In CLARIAH we provide a Dutch version of KALDI for use in context of scholarly research, notably for searching (and automatic transcription of interviews).
Application data	Audio
Application area	Automatic Metadata Generation, Enrichment
Link to software	http://www.opensource-spraakherkenning.nl/index.php/downloads/

Name	Media Suite
Why	In the **media studies** work package the researchers will be supported by integrating improved versions of a range of independently developed applications in one virtual research environment Media Suite. Researchers who want to work with large data collections in Dutch cultural heritage institutions often can not be online with these data because of copyright and privacy reasons. The Media Suite ensures that this is possible.
Description	Media Suite is the online portal that gives researchers access to multimedia data collections in Dutch archives with the help of an authentication and authorization protocol, including tools to work with this data and a workspace to store the various data.
Application data	Multimedia, with special attention to audio, video and images.
Application area	Access
Link to software	mediasuite.clariah.nl

Name	Media Suite Registry
Why	Researchers want an overview of the data collections in Dutch cultural heritage institutions that are available for research. For the underlying infrastructure of the Media Suite, it is important to be able to consult a register of available data collections so that new collections can be added or adjustments / updates to data collections can be implemented.
Description	Media Suite Data Registry is a register based on CKAN for data collections in the Media Suite. The idea is that collection keepers can register their collection metadata based on agreed standards, so that these collections can be made available within the Media Suite for research.
Application data	Collection metadata
Application area	Data
Link to software	http://mediasuite.clariah.nl/data

Name	Media Suite - Collection Inspector
Why	For researchers, it is very important to have a good picture of the quality of the metadata of large data collections. The Collection Inspector provides insight into the underlying metadata and the degree of completeness with the help of visualizations and analytics.
Description	Collection Inspector tool helps researchers to analyze the meaning and quality of collection metadata
Application data	Metadata
Application area	Digital Source Criticism
Link to software	http://mediasuite.clariah.nl/tool/collection-inspector

Name	Media Suite - Search & Select
Why	Researchers are usually interested in parts of large data collections from archives, for example in which a certain subject, person or location is discussed. During the process of distant reading, a researcher looks for relevant content to store it in a so-called personal 'virtual collection' and then start working on the data at a detailed level (close reading). The Media Suite Search & Select tool offers various search options on available collections and the possibility to store and annotate components.
Description	The Media Suite - Search & Select tool offers researchers the possibility to compile personal virtual collections from large data collections at heritage institutions.
Application data	Metadata
Application area	Personal collection building (distant reading)
Link to software	http://mediasuite.clariah.nl/tool/single-search

Name	Media Suite - Compare Tool
Why	Comparing the representation of a subject or person in different collections is something that researchers would like to be able to do, but which is not easy with the standard search options. The Media Suite compare tool makes it possible to generate visualizations for comparing collections.
Description	With the Media Suite - Compare tool the representation of topics in different collections can be visualized.
Application data	Metadata, Queries
Application area	Analysis
Link to software	http://mediasuite.clariah.nl/tool/query-comparison

Name	Media Suite - Explore Tool
Why	In addition to the targeted search in available collections, researchers are interested in a more explorative way of browsing through data collections, combining cross-media information from different (heritage) angles.
Description	With the Media Suite - Explore tool, topics (people, locations, events) can be explored based on Linked Open Data principles on the various (CLARIAH wide) available data collections.
Application data	Metadata, Linked Open Data
Application area	Exploration
Link to software	http://mediasuite.clariah.nl/tool/exploratory-search

Name	Media Suite - Resource Viewer Tool
Why	Researchers working with multimedia data sources have different tools than researchers who mainly work with text. Playback, browsing (no paragraphs or headings) and zooming in with this type of media is different, so also when it comes to annotating it.
Description	The Media Suite - Resource Viewer Tool offers researchers opportunities to get started with the individual multimedia objects in the data collections (or personal virtual collection). It is not only about playing / showing (including zooming in / out), but also about browsing through unstructured objects (where in an hour video it is about a certain subject) using automatically generated segmentations or transcriptions, and to describe it and annotating the media objects (time / space-coded).
Application data	Content, Annotations, Enrichment, Visualisations
Application area	Close reading
Link to software	n.v.t.

Name	Media Suite - Workspace
Why	During the research a researcher generates data (virtual collection, annotations, queries, visualisations) that they want to store and organize neatly, add their own data or export data. The workspace tool is meant for this.
Description	The Media Suite - Workspace tool offers researchers the way to store research data in projects.
Application data	Metadata, Queries, Annotaties
Application area	VRE
Link to software	http://mediasuite.clariah.nl/workspace

Name	Media Suite - Data Analysis & Visualisation Tool
Why	With tools like the Resource Viewer and the Workspace, the Media Suite offers the possibility to do data analysis and make data visualizations. However, researchers also have the need to perform ad-hoc data analysis and visualization for specific research questions.
Description	With the Media Suite - Data Analysis and Visualization Tool, researchers can make data analyzes and visualizations themselves using a "programming interface" (Jupyter Notebooks) on the underlying data infrastructure (APIs).
Application data	Metadata, Annotations
Application area	VRE
Link to software	http://mediasuite.clariah.nl/workspace

Name	PICCL
Why	PICCL offers a workflow for corpus building and builds on a variety of tools. The primary component of PICCL is TICCL; a Text-induced Corpus Clean-up system, which performs spelling correction and OCR post-correction (normalisation of spelling variants etc).
Description	PICCL and TICCL constitute original research by Martin Reynaert (Tilburg University & Radboud University Nijmegen), and is currently developed in the scope of the CLARIAH project. This repository hosts the relevant workflows that constitute PICCL, powered by Nextflow. These will be shipped as part of our LaMachine software distribution. The combination of these enable the PICCL workflow to be portable and scalable; it can be executed accross multiple computing nodes on a high performance cluster such as SGE, LSF, SLURM, PBS, HTCondor, Kubernetes and Amazon AWS. Parallellisation is handled automatically. Consult the Nextflow documentation for details regarding this. PICCL makes extensive use of the FoLiA format, a rich XML-based format for linguistic annotation.
Application data	FoLiA/XML
Application area	OCR, text post-correction and NLP
Link to software	https://github.com/CLARIAH/wp23-PICCL
Comment	PICCL was developed in a joint project funded by CLARIAH WP2 and WP3.

Name	SPOD
Why	The webtool PaQu allows one to upload Dutch texts for automatic parsing. The resulting parsed material can be queried in a number of ways. Making one's own queries for online searching and counting is possible, but time-consuming, especially if you do not use PaQu very often, or if your queries are very complex. For power users this is no problem, for everybody else, there is now SPOD.
Description	SPOD (= Syntactic Profiler of Dutch) is a series of ready-to-use queries which help users to quickly get an overview of a wide variety of syntactic properties. The profiler is part of the PaQu website.
Application data	Data is plain text in UTF-8 or higher encoding.
Application area	Linguistics, stylistics
Link to software	http://haytabo.let.rug.nl:8067/spod

Name	Timbuctoo
Why	Timbuctoo is specifically designed for academic research in the Arts and Humanities, which often yields complex and heterogeneous data. It lives up to academic standards for working with such content: the infrastructure accommodates different views on a subject and leaves the interpretation of the data to the researcher. Also, Timbuctoo keeps meticulous track of data provenance and does not impose a certain research methodology on its users. Data can be searched and analyzed through the web interface, or queried using the API.
Description (a)	Timbuctoo is used to share its data with the world and to host high quality datasets of ongoing research projects. Timbuctoo forms the backbone of Anansi is the central linked open data hub in the CLARIAH infrastructure. Furthermore, Anansi links up with large-scale existing data infrastructures outside CLARIAH and allows researchers to connect their own datasets.

Name	Timbuctoo
Description (b)	The basic structure of the software is a set of REST API's on top of a linked data store (implemented on Berkeley DB), offering developers the opportunity to build clients interacting with these data. Currently the infrastructure provides: - end user GUI's for uploading, configuring and searching a dataset; - the ability to access the data as an RDF graph or as a REST (document oriented) datastore; - various importers for binary formats; - the ability to discover and download datasets from remote servers that contain ResourceSync descriptions; - the ability to subscribe to changes on a dataset, which enables the creation of post-hoc data stores (e.g. MongoDB, MySQL) optimised for specific query patterns. - Timbuctoo is constantly being maintained and updated by the developers of Huygens ING. It is open-source and freely available under GPL v3. Please refer to GitHub for a detailed description of technological features and information for engineers of research institutions who wish to implement in their digital infrastructure.
Application data	CSV, JSON, GraphML, RDF
Application area	Linked Data
Link to software	https://github.com/CLARIAH/timbuctoo

Name	VU-reading-machine
Why	Many texts tell stories. Stories can be seen as sequences of events involving participants. The VU-reading-machine detects events in Dutch texts in terms of what happened, who is involved, where and when. In addition, it determines source perspective on these events: who believes or denies what, what emotion or sentiment is exhibited.
Description	This package is a comprehensive tool (VU-reading-machine v3) for batch processing of raw text files in UTF-8 format. It generates XML output in the Natural Language Annotation Format (NAF) which can be converted to a semantic web format in RDF-SEM and RDF -GRaSP. The RDF output can be semantically queried through SPARQL, e.g. to create sequences of events on timelines.
Application data	Text files in UTF-8 format
Application area	Linked Data: batch processing of text to generate semantic interpretations, such as events, entities, time-expressions, etc. that make up stories
Link to software	https://github.com/cltl/vu-rm-pip3

Name	Yasgui (Yet Another Sparql GUI)
Why	To enhance the efficiency of SPARQL query writing, to share SPARQL queries, and to visualize SPARQL query results as they were supposed to (e.g. geographic output as a map, in addition to a table).
Description	Writing SPARQL queries can be a tedious job, with your usual missing-brackets-hazards and the additional 'what-was-that-prefix-again' sorrows. Yasgui autocompletes prefixes that are stored at generic repositories (e.g. purl), but also provides syntax highlighting, error detection and debugging. Moreover, query results from YASGUI can be directly visualized. Not just as pie charts and graphs, but also image galleries and maps. In addition it provides an API to link to the SPARQL query, allowing queries to be shared and retrieved.
Application data	Any SPARQL endpoint
Application area	Any field using Linked Data
Link to software	yasgui.org
Comment	Yasgui was developed by Laurens Rietveld before the start of CLARIAH, but has been augmented during CLARIAH and has become a central tool in the CLARIAH Linked Data pipeline and the Linked Data workflow in general.

KEY PUBLICATIONS

Aasman, S., Melgar Estrada, L., Slootweg, T. & Wegter, R., (2019). Tales of a Tool Encounter: Exploring Video Annotation for Doing Media History, *VIEW Journal of European Television History and Culture, Special Issue on Audiovisual Data in Digital Humanities, eds. Pelle Snickars, Mark Williams and Andreas Fickers*, Spring 2019. Open access, online multi-media article.

Ashkpour, A., Meroño-Peñuela, A., & Mandemakers, K. (2015). The Aggregate Dutch Historical Censuses: Harmonization and RDF. Historical Methods: A Journal of Quantitative and Interdisciplinary History, 48(4), 230-245. (PDF)

Bilgin, A., Sang, E.T.K., Smeenk, K., Hollink, L., van Ossenbruggen, J., Harbers, F. and Broersma, M. (2018). Utilizing a Transparency-driven Environment toward Trusted Automatic Genre Classification: A Case Study in Journalism History. *Proceedings 14th International Conference on e-Science (e-Science) (pp. 486-496). IEEE*.

Bloothooft, G., Oosterlaken, R., Reynaert, M., Depuydt, K., Schoonheim, T. (2018). NAMES: Towards gold standards for personal names. *DHBenelux conference 2018*.

Broersma, M., & Harbers, F. Eds. (2019). Dossier CLARIAH Media projects. *Tijdschrift voor Mediageschiedenis/Journal for Media History*.

Dijk, J. van (2016). Big data, grand challenges: On digitization and humanities research. *KWALON special issue on Qualitative Research in the Digital Humanities* 2016-1 (61).

Erp, M. van, Does, J. de, Depuydt, K., Lenders, R. & Goethem, T. van (2018). Slicing and Dicing a Newspaper Corpus for Historical Ecology Research. *Proceedings of the 21st International Conference on Knowledge Engineering and Knowledge Management (EKAW 2018)*.

Erp, M. van, Goethem, T. van, Depuydt, K. & Does, J. de (2017). Towards Semantic enrichment of Newspapers: A Historical Ecology use case. *Proceedings of Workshop on Humanities in the Semantic Web – WHiSe II*.

Hagedoorn, B., & Sauer, S. (2019). The Researcher as Storyteller: Using Digital Tools for Search and Storytelling with Audio-Visual (AV) Materials. *VIEW Journal of European Television History and Culture, Special Issue on Audiovisual Data in Digital Humanities, eds. Pelle Snickars, Mark Williams and Andreas Fickers, Spring 2019*. Open access, online multi-media article.

Hoekstra, R.,Meroño-Peñuela, A., Dentler, K., Rijpma, A., Zijdeman, R. & Zandhuis, I. (2016). An ecosystem for Linked Humanities Data. *CEUR Workshop Proceedings*, 85-96.

Hogervorst, S. (2019), Discussiedossier, Inleiding. Oral history en geschiedenisonderwijs. *Tijdschrift voor Geschiedenis* 131 (4), 631-635.

Hogervorst, S. (2018). Distanced by the Screen. Student History Teachers and Video Archives of Second World War Interviews in the Netherlands, *W. Dreier, A. Laumer, M. Wein eds., Interactions. Education with Testimonies*, Vol.4 (Berlin, Stiftung 'Erinnerung, Verantwortung und Zukunft'), 145-153.

Inel, O., Sauer, S. & Aroyo, L. (2018). A study of narrative creation by means of crowds and niches. *A. Bozzon, & M. Venanzi (Eds.), HCOMP 2018 Works in Progress and Demonstration Papers (HCOMP WIP&DEMO 2018): Proceedings of the HCOMP 2018 Works in Progress and Demonstration Papers. Track of the Sixth AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2018). Zurich, Switzerland, July 5-8, 2018.* (pp. 1-4). (CEUR Workshop Proceedings; Vol. 2173).

Lotze, K. (2018). Screening Desmet: reflections on the MIMEHIST-project and CLARIAH's Media Suite. *CLARIAH WP5 blog post, August 27, 2018*, link: https://clariah.github.io/mediasuite-blog/blog/2018/08/27/screening-desmet

Melgar, L., Noordegraaf, J., Koolen, M. & Blom, J. (2017). From Tools to "Recipes": Building a Media Suite within the Dutch Digital Humanities Infrastructure CLARIAH. *In Proceeding of Digital Humanities (Benelux), 3-5 July 2017, Utrecht*.

Melgar, L. & Koolen, M. (2017). Facilitating Fine-grained Open Annotations of Scholarly Sources. *In Proceedings of Digital Humanities 2017, 8-11 August 2017, Montreal, Canada*.

Melgar, L., Koolen, M., Huurdeman, H. & Blom, J. (2017). A Process Model of Scholarly Media Annotation. *Proceedings of ACM SIGIR Conference on Human Information Interaction & Retrieval (CHIIR), 7-11 March 2017, Oslo, Norway.*

Molen, B. van der & Pieters, T. (2017). Distant and close reading of Dutch drug debates in historical newspapers. Possibilities and challenges of big data research in historical public debate research. *Arun K. Somani, Ganesh Chandra Deka (eds.). Big Data Analytics. Tools and Technology for Effective Planning*. Boca Raton: CRC Press: 373-390.

Molen, B. van der, Gorp, J. van & Pieters, T. (2019). Public debates": a methodological operationalization based on development of and reflection on CLARIAH Media Suite tools using digitized heterogeneous mass media datasets. *Clarin accepted publication*

Odijk, J. (2015). Linguistic Research with PaQU. *Computational Linguistics in The Netherlands journal,* 5, 3-14.

Odijk, J. (2016). Linguistic Research in the CLARIN Infrastructure. *Lingua; Vol. 178*. Elsevier.nl.

Odijk, J. (2016). Linguistic research using CLARIN. *Lingua, 178*, *1-4*. DOI: 10.1016/j.lingua.2016.04.003

Odijk, J., & Hessen, A. van (Eds.) (2017). CLARIN in the Low Countries. *London, UK: Ubiquity Press*. DOI:https://doi.org/10.5334/bbi

Olesen, C., & Kisjes, I. (2018). From Text Mining to Visual Classification: Rethinking Computational New Cinema History with Jean Desmet's Digitized Business Archive, *Tijdschrift voor Mediageschiedenis*, Vol. 21, No. 2 (2018): 127-145.

Peursen, W. van (2018). Linked Pasts IV: Views From Inside The LOD-cloud, *Mainz*, 11–13 December 2018. See powerpoint on Slideshare.

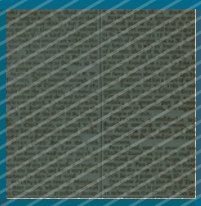
Randeraad, N. (2019). Circulations charitables: 1876-1913), *S. Baciocchi, T. David & C. Topalov (eds.), Philanthropes en 1900 (Londres, New York, Paris, Genève) (Paris, Creaphiseditions, 2019) forthcoming (with Chris Leonards).*

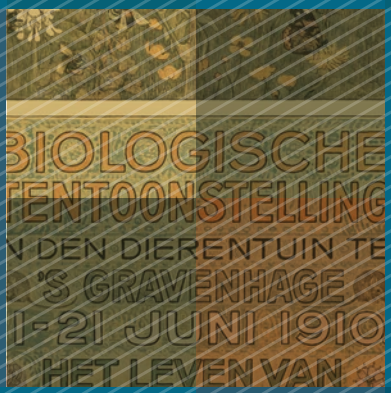
Randeraad, N. (2018). Dutch Social Reformers in Transnational Space, 1840-1914: *Reflections on the CLARIAH Research Pilot 2TBI*, https://www.clariah.nl/projecten/research-pilots/2tbi

Reynaert, M., Gompel, M. van, Sloot, K. van der & Bosch, A. van der (2016). PICCL: Philosophical Integrator of Computational and Corpus Libraries. *Proceedings of CLARIN Annual Conference 2015: Book of Abstracts. De Smedt, K. (ed.)*. Wrocław, Poland: CLARIN ERIC, p. 75-79.

Reynaert, M. (2016). AHA: Anagram Hashing Application. *Proceedings of CLARIN Annual Conference 2016: Book of Abstracts. Borin, L. (ed.)*. Aix-en-Provence, France: CLARIN ERIC.

Wouden, T. van der, Audring, J., Bennis, H. et al. (2016). Het Taalportaal: Een nieuwe wetenschappelijke grammatica voor het Nederlands en het Fries (en het Afrikaans), *Nederlandse Taalkunde*, Vol. 21, No. 1. ISSN 1384-5845.





CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities

NWO
Netherlands Organisation for Scientific Research

WWW.CLARIAH.NL