BMGN - Low Countries Historical Review | Volume 128-4 (2013) | pp. 55-77

Big Data for Global History

The Transformative Promise of Digital Humanities[1]

JORIS VAN EIJNATTEN, TOINE PIETERS
AND JAAP VERHEUL

This article discusses the promises and challenges of digital humanities methodologies for historical inquiry. In order to address the great outstanding question whether big data will re-invigorate macro-history, a number of research projects are described that use cultural text mining to explore big data repositories of digitised newspapers. The advantages of quantitative analysis, visualisation and named entity recognition in both exploration and analysis are illustrated in the study of public debates on drugs, drug trafficking, and drug users in the early twentieth century (WAHSP), the comparative study of discourses about heredity, genetics, and eugenics in Dutch and German newspapers, 1863-1940 (BILAND) and the study of trans-Atlantic discourses (Translantis). While many technological and practical obstacles remain, advantages over traditional hermeneutic methodology are found in heuristics, analytics, quantitative trans-disciplinarity, and reproducibility, offering a quantitative and trans-national perspective on the history of mentalities.

Introduction

What new promises does the rapidly expanding and increasingly varied academic field of digital humanities hold for historians? Humanities scholars are increasingly incorporating computational tools and methods in all phases of their research. Digital tools are used in opening up, presenting and curating textual and multi-media sources, in heuristic techniques of retrieval and accumulation of digitised data, in data analysis, in various forms of visualisation and in enhanced and multi-media publications of research results. In short, digital humanities can be said to touch upon all aspects of

URN:NBN:NL:UI:10-1-110022 | www.bmgn-lchr.nl | e-ISSN 2211-2898 | print ISSN 0615-0505

humanities scholarship. It is not a unified field or methodology, however. It is 'an array of convergent practices' that come together around digitised data and digital tools.[2] An indication of its growing establishment are the large investments in digital infrastructure supported by funding agencies and the emergence of digital humanities centres that aim to support and bring together these efforts worldwide. There are more than one hundred such centres at present, and that number is growing.[3]

Digital humanities has already been called a 'visionary discourse' with a 'utopian core', forwarded by a 'movement', a 'revolution' and even an 'insurgency' within the humanities, complete with its own manifestos, meeting grounds and networks.[4] One recently published volume boldly claims that the field of digital humanities opens up one of the great cultural-historical transformations in human history,

> [...] precisely because it brings the values, representational and interpretive practices, meaning-making strategies, complexities, and ambiguities of being human into every realm of experience and knowledge of the world. It is a global, trans-historical, and trans-media approach to knowledge and meaning-making.[5]

These assertions may sound like histrionics to some historians. Indeed, the application of computational techniques to the humanities is also meeting scepticism and criticism. Literary theorist Stanley Fish recently compared 'the digital vision' to a theology that aims to transcend earthly limitations by once and for all releasing mankind from the confines of 'discrete, partial and

1 The authors would like to thank the editors of BMGN - *Low Countries Historical Review*, the guest editor and the anonymous reviewers for their helpful comments on an earlier version of this article.

2 Peter Lunenfeld, Todd Presner and Jeffrey Schnapp, 'Digital Humanities Manifesto 2.0', 2009, 2; http://hastac.org/node/2182. For an overview of recent trends and discussions in the field of digital humanities see Amy E. Earhart and Andrew Jewell (eds.), *The American Literature Scholar in the Digital Age* (Ann Arbor, MI 2011); Matthew K. Gold (ed.), *Debates in the Digital Humanities* (Minneapolis 2012); http://dhdebates.gc.cuny.edu/debates/part/1; Claire Warwick, Melissa M. Terras and Julianne Nyhan, *Digital Humanities in Practice* (London 2012); Anne Burdick et al. (eds.), *Digital_Humanities* (Cambridge, MA 2012); David M. Berry (ed.), *Understanding Digital Humanities* (Houndmills, etc. 2012).

3 'centerNet | An International Network of Digital Humanities Centers' (10 March 2013), http://digitalhumanities.org/centernet/; Marja Makarow, Milena Zic Fuch and Claudine Moulin, *Research Infrastructures in the Digital Humanities, Science Policy Briefing* (European Science Foundation, 2011); John Unsworth (ed.), *Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences* (American Council of Learned Societies, 2006).

4 Lunenfeld, Presner and Schnapp, 'Digital Humanities Manifesto 2.0'.

5 Burdick et al., *Digital_Humanities*, vii.

situated' knowledge. Some scholars even feel the urge to clarify why so many humanities scholars are reluctant to convert to the new digital creed. Others simply ignore the hype.[6] If digital humanities is a new paradigm, it is one that comes with its stalwart believers and underwhelmed agnostics. We believe that overstatements and emotions will do little to help us assess the intrinsic value of these new digital methodologies for the humanities. It is important to ask ourselves though, whether these new technologies and approaches will change the nature of historical inquiry. Will we see a gradual change, with tools or techniques increasingly being added to established practices, or will digital humanities inaugurate a fundamentally new way of framing historical questions? Are we facing a revolution, as some seem to suggest, comparable to the influx of Byzantine scholarship that led to the re-evaluation of Antiquity on the eve of the Renaissance, or to the opening of national archives at the beginning of the nineteenth century that radically transformed history writing? The question whether the application of computational methods to the humanities holds a transformative promise seems urgent.[7]

One way to address this question is to look at the possibilities offered by the availability of so-called big data in combination with the emergence of digital tools that enable us to mine and analyse gargantuan quantities of digital sources in innovative ways. What is the promise of research projects in which innovative interpretative techniques are applied to big data repositories that are now increasingly being opened up by digital humanities research infrastructures? From a humanities point of view, big data refers to huge quantities of digitised information that can be analysed using data-intensive methods but for which conventional humanist methods, geared as they are towards the interpretation of a limited number of texts, images and data sets, are simply inadequate. It was in this sense that the celebrated *Digital Manifesto 2.0* of 2009 called for a new wave of scholarship that would be 'qualitative, interpretive, experimental, emotive, generative in character' and predicted the emergence of 'bigger pictures out of the tesserae of expert knowledge'.[8]

6 Stanley Fish, 'The Digital Humanities and the Transcending of Mortality', *New York Times* 9 January 2012; http://opinionator.blogs.nytimes.com; Helle Porsdam, 'Too Much "digital", Too Little "Humanities"?: An Attempt to explain why many Humanities Scholars are Reluctant Converts to Digital Humanities' (arcadia@cambridge Publications, 2011), http://arcadiaproject.lib.cam.ac.uk/docs/DigitalHumanities.pdf.

7 Patrik Svensson, 'DHQ: Digital Humanities Quarterly: Envisioning the Digital Humanities', *Digital Humanities Quarterly* 6:1 (17 March 2013), http://www.digitalhumanities.org/dhq/vol/6/1/000112/000112.html.

8 Lunenfeld, Presner and Schnapp, 'Digital Humanities Manifesto 2.0', 2, 4. See for definitions of Big Data in a historical context, the Historicizing Big Data Project of the Max Planck Institute for the History of Science in Berlin, http://www.mpiwg-berlin.mpg.de/en/research/projects/DeptII_Aronova_Oertzen_Sepkoski_Historicizing.

Will big data re-invigorate big history? What large outstanding questions can historians hope to address by implementing digital humanities? In what follows we will gauge some of the new avenues currently being explored in the Netherlands, with a particular focus on the projects WAHSP and BILAND (funded by CLARIN, one of the European research infrastructures), and the project Translantis and its tool Texcavator (funded by the Netherlands Organisation for Scientific Research (NWO)). Both projects revolve particularly around finding new digital forms of the quantitative history of mentalities. Our premise is that new text mining techniques for big data analysis will not replace traditional hermeneutic methods in historical research. Rather, the two should be seen as complementary.

A new turn in digital historical methodology[9]

Most researchers explore data manually, using their knowledge and expertise to extract the information they deem relevant. Media research as a field is almost inherently interested in discovering large patterns of opinion formation. Media historians have traditionally employed a variety of sampling methods. An example from the Netherlands is the methodology that media historian Frank van Vree adopted when he studied Dutch public opinion regarding Germany in the period 1930-1939. This was one of the first studies to use public media to gain insight into what the French have called *histoire des mentalities* – the history of mentalities. Using newspapers as the most important mass medium of the interwar period, Van Vree selected four titles that each represented a major population group (such as Catholics and Protestants). Newspaper issues were then browsed manually, yielding a selection of almost 4,000 articles expressing an opinion on the subject. The 'neutral' press, with a market share of about forty-five per cent, was left out of

9 The most common approach to meet this challenge in historical research is to use statistically grounded sampling methods, such as simple or stratified sampling, snowball sampling, and sampling with replacement. Simple sampling refers to the random selection of individual data from a single population, a method that is sometimes refined by sampling from sub-populations or strata; snowball sampling uses a small selection of initial data to select further data (comparable to the way social networks expand through the selection of 'friends'); while sampling with replacement is a form of random sampling that leaves open the possibility that individual data are selected more than once. These and other methods are well described in the humanities literature and have often been used in content analysis of newspapers as well as magazines. See for instance: J. Gary Knowles, *Handbook of the Arts in Qualitative Research: Perspectives, Methodologies, Examples, and Issues* (Los Angeles 2008); Klaus Bruhn Jensen, *The Handbook of Media and Communication Research: Qualitative and Quantitative Methodologies* (New York 2012).

consideration. More recently, Els Witte followed a similar approach in a study on the image of the nation in the Belgian Revolution. Six newspapers with varying political signatures from different cities were selected and browsed manually, yielding 350 articles that expressed an opinion.[10] These studies are successful examples of a wide, international field of inquiry into the role of mass media in historical discourse that can be said to have opened up new perspectives on the study of 'public opinion' or collective 'mentalities'.

Where big data is involved however, it is impossible to analyse all relevant articles by browsing, while making a selection through a sampling method becomes increasingly problematic because the end selection always needs to be manageable for an individual researcher. Indeed, historians have recognised since at least the 1970s that there are corpora available for historical research that are simply too large to be examined in their entirety and to be perused manually. Nevertheless, manual browsing is still common practice in historical research. Conventional sampling methods to some extent do address the challenge of big data in that they reduce the amount of data to manageable proportions, but in practice they are relevant only to the analysis of a limited number of serial titles.[11]

Text mining big data collections: program design

One of the expectations about digital humanities is that it will enable us to investigate much larger quantities of public media. After half a century of digital humanities we are now entering a new phase in which historians are able to analyse massive volumes of texts, particularly by integrating (socio-) linguistic methods into humanities research. New techniques of large-scale data analysis allow historians to manage big data sets that were difficult to access earlier. Semantic text analysis is a particularly promising form of data mining that can be applied to textual data in order to derive subject-specific information from 'mountains' of textual data without having to read it all. Text analytics or text mining is an umbrella term for incorporating and implementing a wide range of tools or techniques (algorithms, methods),

10 Frank van Vree, *De Nederlandse pers en Duitsland, 1930-1939. Een studie over de vorming van de publieke opinie* (Groningen 1989); Els Witte, *De constructie van België, 1828-1847* (Leuven 2006). For American examples of sampling to chart public debates in magazines see Celeste Michelle Condit, *The Meanings of the Gene: Public Debates about Human Heredity, Rhetoric of the Human Sciences* (Madison 1999).

11 Recent developments in media history are discussed in Adrian Bingham, 'Reading Newspapers: Cultural Histories of the Popular Press in Britain', *History Compass* 10 (n.d.) 140-150; Roderick P. Hart and Elvin T. Lim, 'Tracking the Language of Space and Time, 1948-2008', *Journal of Contemporary History* 46 (2011) 591-609.

including data mining, machine learning, natural language processing and artificial intelligence. Semantic text analytics focuses specifically on the historical-contextual meanings of words and phrases in a big data set.[12]

The goal of text mining is to reduce the effort required of humanities researchers to obtain useful information from large digitised textual data sources. Current international and national programmes such as Digging into Data and CATCH-plus demonstrate the feasibility of performing interdisciplinary humanities research facilitated by digital research tools.[13] Adapting the digital methodologies arising from these programmes to humanities research gives rise to more easily reproducible results, more refined computationally-based research methods for historians and new research questions. These programmes also demonstrate that collaborative and integrative strategies such as common group learning (all knowledge is necessary pooled and learning is both shared and cumulative), modelling, negotiation among experts and integration by leaders are central to the functioning and therefore the success of this approach. The design and execution of such large digital humanities programmes is obviously grafted on common practice in the sciences and may be contrasted to the great majority of humanities research (exceptions excluded notably linguistics), where research is predominantly individualistic.

The role of experts in the field, in our case cultural and science historians, in the development of new text and data mining technologies is particularly important. The process of articulating the needs and demands of users in relation to available technical options is no less significant and crucially depends on including programme mediators who bring a strong background in the humanities as well as state-of-the-art text mining expertise into the research team. Incorporating regular feedback loops for instance, allows an iterative refinement of analysis algorithms and the development of a user-friendly digital tool. In the following sections we will illustrate two particular programs, WAHSP/BILAND and Translantis.

Towards historical sentiment mining in public media: WAHSP/BILAND

The first step towards the development of an open-source mining technology that can be used by historians without specific computer skills is to obtain a hands-on experience with research groups that use currently available

12 Peter Jackson and Isabelle Moulinier, *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization* (Second revised edition; Amsterdam 2007).

13 www.diggingintodata.org; www.catchplus.nl.

open-source mining tools. A recently developed tool that has been utilised to accomplish this is the CLARIN-supported web application for historical sentiment mining (a form of semantic text analytics that focuses on historical opinions, attitudes, and value judgments) in public media that is known under its acronym WAHSP.[14] WAHSP is specifically designed for text mining the digital newspaper archive of the National Library of the Netherlands (*Koninklijke Bibliotheek*). At present, this repository comprises over eight million pages from more than 200 different newspapers and periodicals published between 1618 and 1995, all together about 100 million articles. The technical basis of WAHSP is the open-source software infrastructure xTAS, which has been developed by the Intelligent Systems Lab at the University of Amsterdam (ISLA).[15] This open-source platform for text analytics has also been applied and tested in computational humanities projects such as Dutch Language Online Media Analysis (STEVIN), Building Rich Links to Enable Television History Research-BRIDGE (CATCH), Elite Network Shifts (KNAW), Infiniti (COMMIT), and Political Mashup.[16]

Although the WAHSP-tool offers a number of options for quantitative analysis, such as the frequency of words or combinations of words used in specific newspaper articles in a certain period of time, it derives its most promising analytical potential from its visualisation and arrangement features. Each query results in a term cloud that is based on the relative frequencies of the words occurring in the retrieved selection of documents from the corpus. The visualisation of word associations in these term clouds allows the historian, on the basis of existing domain expertise, to quickly determine the characteristics of the selected documents and to refine or adapt the query. The WAHSP software is also able to indicate sentiments by highlighting terms with a negative or positive connotation (although it should be noted that this technique of sentiment detection is still in need of historical contextualisation). Advanced techniques for what is called Named Entity Recognition (NER) enable the researcher to recognise and highlight the names of 'entities' such as places, persons, institutions and events. This tool allows the historian to place the occurrence of certain terms, ideas or debates within a geographical context, or connect them to persons or organisations (see fig. 3).

Lastly, a visualisation of the temporal distribution of the documents allows the historian to discover patterns in publication dates. This visualisation is a histogram plot of publication dates that can be explored

14 http://www.clarin.nl/page/about/2; http://biland.nl.

15 http://xTAS.science.uva.nl.

16 http://ilps.science.uva.nl/biblio/duoman-dutch-language-online-media-analysis; http://ilps.science.uva.nl/node/735; http://ilps.science.uva.nl/news/knaw-computational-humanities-grant; http://ilps.science.uva.nl/research/projects/bridge; http://www.kitlv.nl/home/Projects?id=25; http://www.project-infiniti.nl/; http://politicalmashup.nl/.

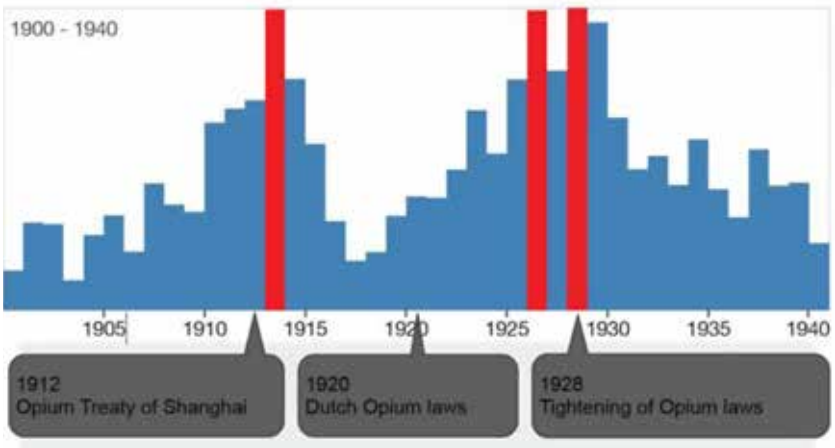
interactively (fig. 1); zooming in on specific parts of the histogram provides finer-grained data (fig. 4). To enable quick recognition of atypical patterns, bursts within the histogram – time periods where significantly more documents were published in comparison to periods around that burst – are highlighted (fig. 1). Clicking on a burst yields a visualisation of word associations of that burst alone and a list of documents contained within that burst (fig. 5). At any point in time the historian can return to the original text of the newspaper article. This allows the historian to get an in-depth understanding of what each burst is about. Together these interactions allow historians to interactively investigate the document selection in order to detect patterns, improving the representativeness of the selection.[17]

The main added value of this digital tool lies in its possibilities for exploratory reading of historical patterns in public debates. The WAHSP research team found that, in terms of methodology, semi-automatic document selection fits rather well with historical research as an alternative to manual browsing or random sampling, facilitating the combination of qualitative and quantitative approaches. Through text mining and visualisation, new insights can be gained from an initial selection. Word clouds depicting the linguistic context within which keywords occur are instrumental in helping the historian with expert knowledge of the domain to combine and compare different historical periods in a free associative manner on the basis of a large number of historical documents. Each query immediately yields a document selection without laborious sampling. This speeds up the heuristic process considerably. Exploring word associations and metadata, as well as visualisations of the documents over time, can lead to improved queries and therefore to a more representative document selection. Such quantitative analysis enhances the knowledge of the historian. A clear benefit of using exploratory searches is to allow the historian to recycle previous insights to investigate new research questions. Comparing document selections using quantitative analysis helps to validate these selections, making them less arbitrary and thus more representative.

This approach has been successfully employed by Stephen Snelders in a WAHSP-assisted study of public debates on drugs, drug trafficking and drug users in the early twentieth century (1900-1940). Building on his domain knowledge, Snelders first marked key events such as the Shanghai Opium Conference (1909) and succeeding treaties, and the introduction and subsequent tightening of the Dutch Opium laws (1920 and 1928) (fig. 1).To gain a better understanding of the dynamics of debates associated with these events, Snelders then determined which terms were associated with drugs like opium, morphine, heroin and cocaine. The comparison of word clouds derived from several periods indicated a shift from a vocabulary related to health issues

17 http://dev.wahsp.nl.

(therapeutic drugs [*geneesmiddelen*], poisons [*vergiften*], addiction [*verslaving*], illness [*ziekte*], pharmacies [*apotheken*]) to one associated with crime-related issues (police [*politie*], smuggling trade [*smokkelhandel*], arrested [*gearresteerd*]), China and war [*oorlog*]). By carefully inspecting the word counts, Snelders found quantitative evidence for historical turning points that indicated the criminalisation of the drugs debate around 1924.[18] Moreover, the WAHSP text mining exercise was instrumental in showing that opium was subject to a politics of 'othering' and that it had a direct bearing on perceptions of China and on the opium distribution and control regime of the colonial state ('Opiumregie') in the Dutch East Indies. At the same time, as earlier studies already concluded, illicit trade is referred to in terms of transit trade through Dutch harbours into other countries.[19]



▲

Fig. 1: Plot of WAHSP search results based on KB newspaper repository using query opium for 1900-1940, courtesy of Daan Odijk (ISLA, UvA; 11-01-2012). The 'bursts' are indicated in red.

18 D. Odijk et al., 'Semantic Document Selection' (presented at the Theory and Practice of Digital Libraries, Paphos (Cyprus); Berlin 2012).

19 Stephen Snelders and Toine Pieters, 'The Blue Lotus Revisited: Public Perceptions of Drug Use in the Dutch Empire, c. 1900-1942', in: *Drugs and Drink in Asia: New Perspectives from History: Proceedings of the Conference of June 22-23, 2012* (Shanghai n.d.); Marcel de Kort, *Tussen patiënt en delinquent. Geschiedenis van het Nederlandse drugsbeleid* (Hilversum 1995).

As a follow-up of the WAHSP-project the bilingual text mining tool BILAND is currently being developed as an open-source and accessible web application. An interdisciplinary team of researchers from the Descartes Centre for the History and Philosophy of the Sciences and the Humanities at Utrecht University and computational scientists from the Intelligent Systems Lab Amsterdam (ISLA) are tailoring WAHSP to the language-specific needs of comparative historical research, with a particular focus on the identity, intensity and location of discourses about heredity, genetics and eugenics in Dutch and German newspapers between 1863 and 1940. The challenge is to incorporate the semantics of two different languages (in this case Dutch and German) and scripts (such as Latin and Gothic). As in WAHSP, BILAND employs a user-oriented, iterative model of collaboration between humanities scholars and ICT developers. Every developmental task and research activity envisaged within the project is a transdisciplinary co-production. This includes selecting and filtering out meaningful lexical items, carrying out text mining tasks, training the algorithms, and meeting the needs of the domain users by including feed-back loops.

Historian Pim Huijnen, employed the BILAND-tool to compare debates about eugenics in Germany and the Netherlands. His goal was to analyse to what extent eugenics debates in these different nations reflected social and cultural notions of individual in relation to collective identities within the context of modernity.[20] Huijnen started to focus on the multiple discourses that converged around the use and adaptation of genetic knowledge and eugenics in the workplace, the home and the wider world.[21] The challenge was to qualify and quantify these 'hidden debates'. In this kind of cultural history (or 'history of mentalities'), the combination of scientific concepts and cultural notions is of primary interest. Thus, we will not only be able to mine concepts but also explore the more unconscious, latent use of genetic or eugenic ideas by ordinary people as they were mediated in public debates.[22]

Using BILAND, Huijnen started with single queries of a fairly straightforward nature including basic concepts such as eugenics and inheritance (see fig. 2, 3 and 4). The BILAND tool demonstrated, in line with the extensive literature on the history of eugenics, that the biological and medical connotation of inheritance was dominant from the end of the nineteenth

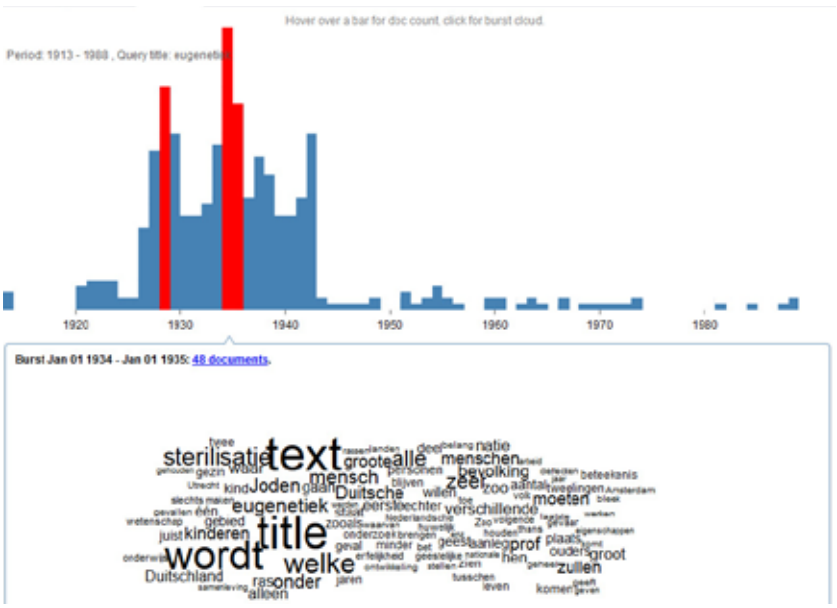
20 Marius Turda, *Modernism and Eugenics* (New York 2010) 124.

21 Stephen Snelders and Toine Pieters, 'Van degeneratie tot individuele gezondheidsopties. Het maatschappelijk gebruik van erfelijkheidsconcepten in de twintigste eeuw', *Gewina* 26:4 (2003) 203-215.

22 Stephen Snelders first elaborated this thesis in Stephen Snelders, 'The Plot against Cancer: Heredity and Cancer in German and Dutch Medicine, 1933-1945', *Gesnerus* 65 (2008) 42-55.

century and throughout the first half of the twentieth century.[23] However, the context in which the concept was debated did change considerably over time. The word cloud for example, makes it plain that the articles containing the word in 1867 predominantly focused on medical issues (fig. 4). In 1935, however, the medical context of using the term inheritance made way for a legal and racial context (fig. 5).

Subsequently, Huijnen searched for combinations of keywords that do not necessarily refer directly to eugenics as such, but which do imply eugenic thinking. Keywords included 'ancestry', 'lineage', 'descent', 'stock', 'reproduction', 'regulation', 'selection', 'pure'/'purity', 'progression', 'evolution', 'deterioration', 'depravation', 'isolation' and 'segregation'. These were entered in combination with keywords from various social or cultural domains such as sports, the home and education. In a similar fashion to the WAHSP-tool, the BILAND-tool helped Huijnen to find an intricate structure of notions concerning heredity that inspired public debates around social policies in Dutch newspapers.



▲

Fig 2: Plot of BILAND search result plus additional word cloud based on KB newspaper repository using query *eugenetiek* ('eugenics') for 1900-1940, courtesy of José de Kruif (UU; 5-12-2013). The 'bursts' are indicated in red.

23 Jan Noordman, *Om de kwaliteit van het nageslacht. Eugenetica in Nederland, 1900-1950* (Nijmegen 1989); Snelders and Pieters, 'Van degeneratie tot individuele gezondheidsopties'.

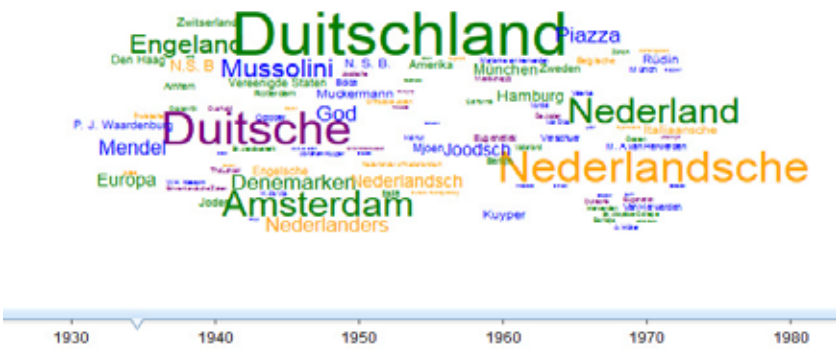


Fig. 3: BILAND search result with named entity recognition cloud based on KB newspaper repository using query *eugenetiek* ('eugenics') for 1900-1940, courtesy of José de Kruif (UU; 5-12-2013).

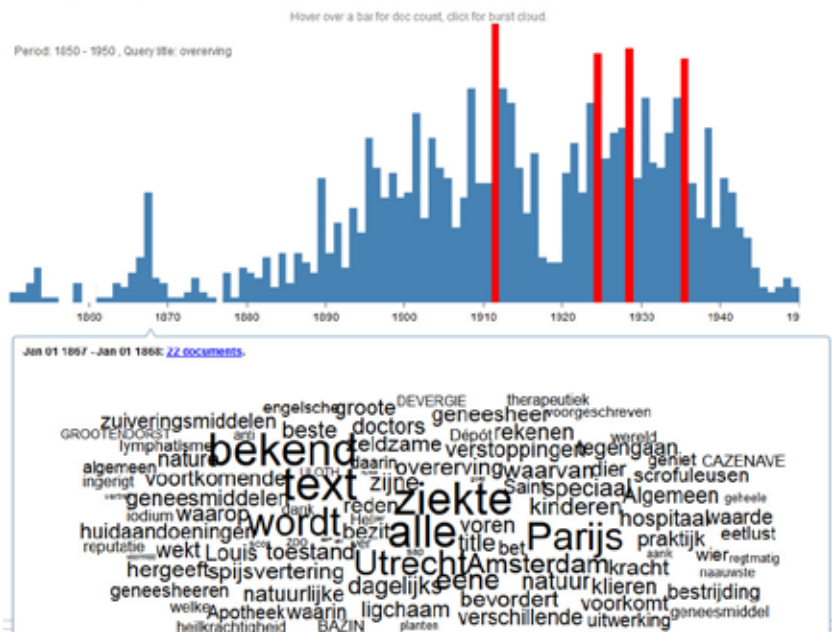
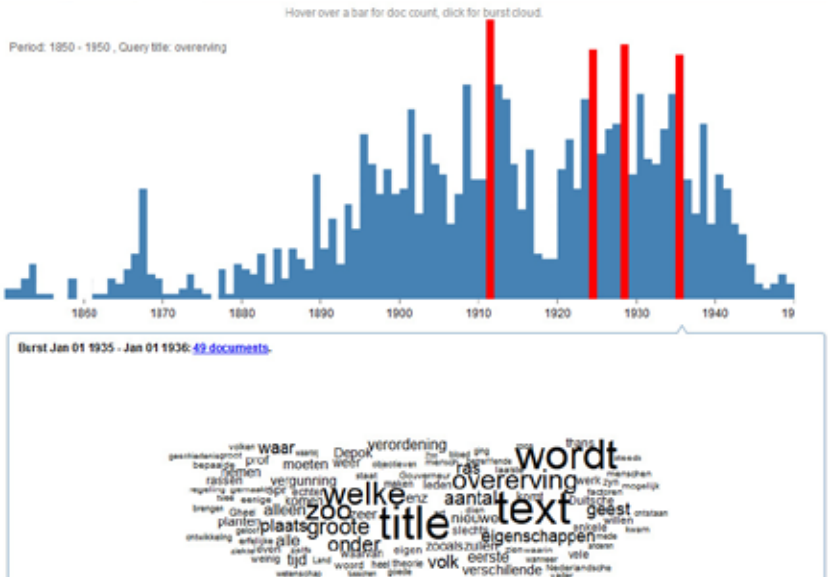
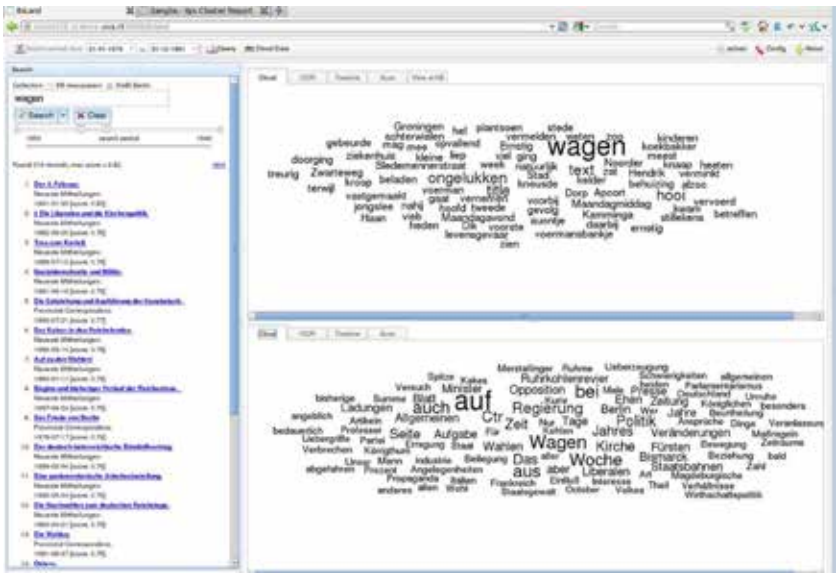


Fig. 4: Plot of BILAND search result plus additional 1867 word cloud based on KB newspaper repository using query *overerving* ('inheritance') for 1850-1990, courtesy of José de Kruif (UU; 5-12-2013). The 'bursts' are indicated in red.



▲

Fig. 5: Plot of BILAND search result plus additional 1935 word cloud based on KB newspaper repository using query *overerving* ('inheritance') for 1850-1948, courtesy of José de Kruif (UU; 5-12-2013). The 'bursts' are indicated in red.



▲

Fig. 6: BILAND search results based on KB and *Staatsbibliothek zu Berlin* digital newspaper repositories using query *wagon* ('wagen') for 1878-1891, courtesy of Fons Laan (ISLA, UvA; 04-09-2013).

Government housing strategies and wage governance for example, were associated with heredity issues. Various combinations of indicators of hereditary thinking ('parentage', 'ancestry', 'innate', 'posterity', 'procreate' and so on[24]) and words related to housing and wage policies ('housing', 're-education', minimum-wage[25]) yielded a significant number of relevant hits, making this particular case worthwhile to look into more carefully, both within the Dutch and German contexts.[26]

The digital methodologies applied in WAHSP and BILAND yielded a number of advantages over traditional research. First of all, the software offered comprehensive overviews of 'bursts' and continuities in debates, and enabled researchers to mobilise research data based on those indicators without time-consuming perusal of all articles. In addition, and more importantly, quantitative – and reproducible – precision could be added to our knowledge on the one hand, about the shift from medical to criminal framing of Dutch drug discourse in the interwar period, and on the other, the role of hidden debates in framing eugenics within the Dutch and German contexts. Thirdly, the word clouds proved a valuable heuristic tool that helped to incorporate into the research project connections that were initially unclear or unsuspected.

A next step in the context of the WAHSP/BILAND projects is to develop tools for cluster analysis (grouping data in particular ways) and multilingual settings (allowing comparisons to be made across languages (for an example, see fig. 6)), and to perform analyses on a much deeper semantic level with the help of historical dynamic lexicons. Above all, multiple research teams should be brought together to exchange experiences in text and data mining and to share research results with other digital humanities scholars, both nationally and internationally.

24 In Dutch: 'afkomst', 'aangeboren', 'voorouders', 'nageslacht', 'voortplanten'.

25 In Dutch: 'huisvesten', 'heropvoeden'.

26 Huijnen is currently in the process of doing so: http://www.biland.nl. P. Huijnen, F. Laan. M. de Rijke, T.A. Pieters, *A Digital Humanities Approach to the History of Science: Eugenics revisited in Hidden Debates by Means of Semantic Text Mining: Histoinformatics* (New York 2013).

Trans-national text mining: Translantis

These new techniques do not only enable us to analyse public debates by tracing the use over time of specific concepts or notions and even 'hidden' debates about specific issues. They also open up a much wider panorama. Text mining techniques offer an innovative way to map trans-national influences and measure how debates crossed regional, cultural and national borders. The availability of massive repositories of digitised periodicals that span not years or decades but centuries, and represent national 'climates of opinion', offers the possibility to map long-term changes in national and trans-national debates on a myriad of issues in their cultural, economic, political and social contexts. It will be evident that there is a link here with global or world history, which as an established sub-discipline is primarily focused on both long-term developments and large-scale comparisons.[27] Given that the larger part of global history research has been exclusively socio-political or economic in nature, one of the most promising new approaches is a study of the interactions between large culturally defined areas such as health, crime, religion and mass communication. The emergence of trans-national power constellations or empires, as well as the way their influence was radiated beyond national borders, can now be mapped over longer periods of time. A way to conceptualise these trans-national and trans-cultural vectors of influence is to investigate the emergence of 'reference cultures'.

The formal title of Translantis is 'E-Humanity Approaches to Reference Cultures: The Emergence of the United States in Public Discourse in the Netherlands, 1890-1990'. The programme defines reference cultures as spatially and temporally identifiable cultures that offer a model to other cultures and have exerted a profound influence in history. Examples of powerful reference cultures are the seventeenth-century Dutch Republic, nineteenth-century Britain and France, early twentieth-century Germany and the United States in the twentieth century – and perhaps twenty-first-century China. Within specific historical periods, each of these major states provided a frame of reference to territories within, or bordering on, their economic, political, technological or military spheres of influence. The impact of reference cultures is therefore pre-eminently trans-national. In this respect, reference cultures form an outstanding research object for digital humanities. They have played a crucial role in the dynamics of global history and involve complex cultural encounters between innumerable and unidentified actors with a variety of perspectives and interests that escape the span of traditional research methods. In contrast to an essentialist and/or territorial concept such as the nation, that of reference cultures allows researchers to address the

27 See for instance the classics Jared M. Diamond, *Guns, Germs, and Steel: The Fates of Human Societies* (New York 1999); Hans Zinsser, *Rats, Lice, and History* (New Brunswick 2008).

shifting subjectivities central to cultural encounters and question rather than assume national identity formation. Reference cultures are mental constructs or 'cognitive maps' that do not necessarily represent a geopolitical reality with an internal hierarchy and recognisable borders. These culturally conditioned images of trans-national models are typically established and negotiated in public discourse over a long period of time.[28]

The academic discussion suggests that the interplay of political, economic and technological supremacy with the 'soft power' of cultural attraction and reputation plays a crucial role in how dominant nations and cultures establish guiding standards for other cultures. However, the specific historical dynamics of reference cultures have never been systematically analysed and hence are not fully understood.

The key to understanding the emergence and dominance of reference cultures is to chart the public discourses in which these collective frames of reference, particularly in the interrelated domains of economy, culture and science and technology are established. The availability of an increasing number of large digital data collections in the public domain, such as the National Library of the Netherlands (KB), the British Library and the US Library of Congress, as well as in commercially available packages enable us – for the first time – to do exactly this: to study long-term developments and transformations in national discourses in a systematic, longitudinal, comparative and quantifiable way. Digital research tools allow us to test the value of reference cultures as a qualitative heuristic historical model and to pair it in a meaningful fashion with quantitative methodology. Essential to the Translantis project is the integration of distant reading of printed media (for instance in the form of a software-enabled assessment of patterns, trends and sentiments in huge amounts of texts) with close reading and interpretation of cultural texts culled from these repositories and from other sources of information within their specific historical context. The study of reference cultures is not only inherently important; it is also instrumental in exploring and identifying the

28 The concept of reference cultures is being developed in the Translantis project. See translantis.nl for further publications on this topic. Previously, such interactions between cultures are mostly discussed in terms of empire. See for instance Charles S. Maier, *Among Empires American Ascendancy and Its Predecessors* (Cambridge, Mass., London 2007); Jane Burbank and Frederick Cooper, *Empires in World History: Power and the Politics of Difference* (Princeton, NJ 2010). The role of public opinion in the establishment of empires is discussed in Catherine Hall, *Cultures of Empire: Colonizers in Britain and the Empire in the Nineteenth and Twentieth Centuries: A Reader* (Manchester, UK 2000); John M. Mackenzie (ed.), *European Empires and the People: Popular Responses to Imperialism in France, Britain, the Netherlands, Belgium, Germany and Italy* (Manchester 2011). See for mental maps the classic study: Alan K. Henrikson, 'The Geographical "Mental Maps" of American Foreign Policy Makers', *International Political Science Review* 1:4 (1980) 495-530.

transformative qualities of digital humanities. New vistas can be opened on the past, including comparative history and the history of mentalities.

As the programme's name suggests, Translantis involves trans-national relations across the Atlantic, in particular those between the Netherlands and the United States. Several considerations make the United States the ideal case to study the underlying twentieth-century dynamics of the use of reference cultures in public debates about economy, culture and science and technology in the Netherlands. As publisher Henry Luce suggested in 1941, the twentieth century can, to a significant degree, be called 'an American Century'. Many scholars have argued that what the English journalist William T. Stead in 1902 presciently dubbed *The Americanisation of the World* has been the result of a deliberate policy of public diplomacy or 'cultural imperialism'. Assuming a direct connection between economic and cultural dominance, historians claimed that the 'coca-colanisation' or 'McDonaldisation' of the world created an 'irresistible empire' of American mass culture, mass consumer goods and consumerism. More recent scholarship has put the effectiveness of these push factors into perspective by underlining the agency of receiving audiences in more complex and reciprocal processes of selective appropriation, 'creolisation' and resistance. If the United States has become a leading culture that has dominated public discourse in European nations, its ascendancy was not the result of a continuous and linear process of cultural transfer, but rather the outcome of numerous distinctive transatlantic encounters in the interrelated domains of economy, culture and science and technology, each with its own participants, dynamics and pace.[29]

The Netherlands seems an especially valuable case study for analyses of discourses about the United States as a means of viewing Dutch society from the outside in. A relatively small nation with a maritime tradition, placed at the border of continental Europe, the Netherlands has developed an openness to foreign ideas and a strong tradition of inward and outward foreign investment. For American-based multinational businesses for example, the Netherlands has long served as a testing ground for Europe. Opinion surveys place the Netherlands at the top of nations most favourable towards the United States within Europe and English is widely spoken. These factors

29 Richard Kuisel, *Seducing the French: The Dilemma of Americanization* (Berkeley 1993); Richard Pells, *Not Like Us: How Europeans have loved, hated, and transformed American Culture since World War II* (New York 1997); Rob Kroes and Robert W. Rydell, *Buffalo Bill in Bologna: The Americanization of the World, 1869-1922* (Chicago 2005); Reinhold Wagnleitner, *Coca-Colonization and the Cold War: The Cultural Mission of the US in Austria after the Second World War* (Chapel Hill 2000); Victoria De Grazia, *Irresistible Empire America's Advance through Twentieth-Century Europe* (Cambridge, Mass. 2005); David W. Ellwood, *The Shock of America: Europe and the Challenge of the Century* (Oxford 2012); Mary Nolan, *The Transatlantic Century: Europe and America, 1890-2010* (Cambridge 2012).

turned the Netherlands into a knowledge economy that became a gateway to Europe for American goods and ideas.

An abundance of recent scholarly work confirms that Luce's American Century took firm hold in the Netherlands. The research programme Dutch Culture in a European Context, for instance, concluded that the United States became the dominant reference culture for the 'contested modernisation' of the Netherlands at the mid-twentieth century, replacing England, France and Germany, countries that had provided the European benchmark for Dutch culture at the beginning of the century. That trans-Atlantic component in the modernisation of the Netherlands has been underlined in major programmes resulting in multivolume overviews of twentieth-century technology and business, such as 'Technology in the Netherlands in the Twentieth Century' (TIN 20, completed 2003) and 'Dutch Business in the Twentieth Century' (BINT). The importance of transatlantic connections is also indicated by the project *Four Centuries of Dutch-American Relations* (completed 2009).[30] Yet, we know remarkably little about the dynamics behind the emergence of the United States as a predominant but contested reference culture for the Netherlands in the twentieth century. What networks or persons and institutions played a role, what was the pace of change and how did the transformative power of the United States as a reference culture change in Dutch public discourse?

Cultural and science historians, information scientists and text mining experts are currently addressing these questions in Translantis.[31] The programme will implement the text mining tools that have emerged from the WAHSP and BILAND projects to study long-term developments and transformations in national discourse in a systematic, longitudinal, and quantifiable way. It is expected that the implementation of text mining tools will provide historians with a sophisticated heuristic model outlining the emergence, role and decline of reference cultures such as the United States, and possibly rising economic powers such as China. The outcomes of this project – insight into reference cultures and the experience with digital technology

30 Jan Bank and Maarten van Buuren, *1900: The Age of Bourgeois Culture, Dutch Culture in a European Perspective* (Basingstoke 2004); Kees Schuyt and Ed Taverne, *1950: Prosperity and Welfare, Dutch Culture in a European Perspective* (Basingstoke 2004); Hans Krabbendam, Cornelis A. van Minnen and Giles Scott-Smith, *Four Centuries of Dutch-American Relations* (Amsterdam, Albany, NY 2009); Johan Schot, Arie Rip and Harry Lintsen, *Technology and the Making of the Netherlands: The Age of Contested Modernization, 1890-1970* (Zutphen 2010).

31 This programme is generously supported by a Horizon research grant of €2 million from the Netherlands Organisation for Scientific Research (NWO) and has started 1 January 2013. Participants include researchers from Utrecht University, the University of Amsterdam, the National Library of the Netherlands (KB) and the Huygens Institute for the History of the Netherlands; the project is led by the authors of the present article.

to mine public debate – will serve as a springboard for comparative studies on European, trans-Atlantic and global levels to determine the patterns of transnational discourse and global cultural exchange.

The digital promise

Based on the examples discussed, we argue that the application of new digital techniques offers a number of methodological advantages over more conventional approaches in humanities research, particularly the history of mentalities. These advantages manifest themselves especially in, but are not necessarily limited to, research that involves textual big data repositories. The advantages arguably apply to at least four related areas – heuristics, analytics, quantitative trans-disciplinarity and new forms of reproducibility.

1. Heuristics

Digital search tools allow searches into textual data of virtually unlimited size, meaning that they are constrained only by the availability of digitised data repositories and computational capacity. This crucial dimension of big data research has several important implications. Firstly, it means that both manual browsing, which is inherently limited in scope, and sampling in its various forms, which involves restrictions with regard to representativeness, are no longer necessary. Secondly, searches no longer depend on indexing or registers, allowing explorations in both structured and unstructured data sets. Thirdly, digital search techniques allow unlimited combinations of searches, facilitating associative thought in a controllable and reproducible way. This fosters creativity and serendipity, bringing the domain specific expertise of the researcher into full play. Even more promising is the potential to discover and quantify 'hidden' debates that offer a new perspective on the history of mentalities.[32]

2. Analytics

Computational techniques allow new ways to analyse research results. The computational methods which are currently being explored in WAHSP, BILAND and Texcavator allow the recognition and display of patterns in textual data. They recognise and identify entities in texts, such as proper names, events and geographical locations, and reveal historical arrays of sentiments and values. Combination of these data and the use of metadata (data that describe the

32 For a discussion about the potential of digital humanities to unlock hidden information and developing new knowledge, see also Diane McDonald and Ursula Kelly, 'Value and Benefits of Text Mining', *IISC*, 14 March 2011, http://www.jisc.ac.uk/publications/reports/2012/value-and-benefits-of-text-mining.aspx.

structure or content of data in a repository) allow the researcher to reconstruct the structure, intensity and emotions of historical debates in public media. Statistical forms of cluster analysis can point to patterns in debates that possibly eluded the traditional researcher. This offers novel ways to trace the course of debates in newspapers and other public media, hence offering a new perspective on public debates.

3. Quantitative trans-disciplinarity

Big data analytics not only opens up new fields of research in cultural history, but also results in quantitative data sets that can be confronted with other historical data. Especially promising are combinations between quantitative textual data and historical data sets that are being produced by economic and social historians, and data that have been accumulated by national, European and global institutions such as Statistics Netherlands, Eurostat, OECD historical statistics, UN Data and WTO statistics.[33] Translantis explores national data sets, connecting local and global historical developments.[34] Mining textual big data opens new vistas on global history, which is still largely dominated by hypotheses proposed by quantitative social and economic research. These can now be tested and contested by reproducible textual data related to cultural interconnections, patterns of value transfer and global *histoires croisées*.

4. Reproducibility

Although reproducibility has always been an ambition in some quarters of historical scholarship, digital approaches now make it possible to publicly document and trace all stages of the historical research process. Enriched publications allow tracing of the heuristic search and selection process, the data sets and open-source software that were used in analytics, and in the statistical underpinnings of transdisciplinary comparisons and contrasts.

In brief, we suggest that the digital innovations discussed in this paper make possible a truly new form of quantitative history of mentalities that avoids some of the pitfalls of conventional research. Given the increasing availability of digitised material, text mining techniques will be indispensable regardless of methodological discussions about the pros and

33 www.cbs.nl/; epp.eurostat.ec.europa.eu/; http://www.oecd-ilibrary.org/statistics; http://data.un.org/; http://www.wto.org/english/res_e/statis_e/statis_e.htm. For an example of how such statistics information can be used to analyze transatlantic differences and similarities, see Peter Baldwin, *The Narcissism of Minor Differences: How America and Europe are Alike* (New York 2011).

34 Another programme, which will start in the course of 2013 and which has not been discussed in this paper is the HERA-funded project Asymmetrical Encounters: Digital Humanities Approaches to Reference Cultures in Europe, 1815-1992 (Asymenc). This program builds on BILAND and Translantis, focusing especially on the trans-national dimension.

cons of conventional sampling: not only will more historical sources be made accessible retrospectively in digital form, but the computerised documentation of today will be the historian's material of tomorrow.

Concluding remarks

We expect that methods geared to text mining big data will set agendas for historical research, in the sense that they will determine what the significant themes in public debates within a specific time frame actually were. This kind of fingerprinting or fixation of historical mentalities on the basis of big data evidence is unprecedented. It might well – at long last – allow humanities scholars to validate their inferences in the manner of social scientists. However, it remains to be seen whether the conclusions obtained through digital methodologies will differ substantially from those acquired by traditional means. Ascertaining this is one of the aims of the Translantis project outlined above and also of the European follow-up HERA project 'Asymmetrical Encounters: E-Humanity Approaches to Reference Cultures in Europe, 1815-1992 (ASYMENC)' that started in the autumn of 2013.

Convinced of their transformative power as we may be, we also readily recognise there are downsides to big data research as proposed here. Given the scarcity of research money, financial investments in quantitative big data research will inevitably occur at the expense of 'traditional' humanities research. We believe that the kind of 'digital mentalities' research outlined in this article will make it possible to quantify the cultural baggage of groups of people – to identify changes in their sentiments, attitudes and values over decades, if not centuries. Yet to interpret such changes the hermeneutic skills traditionally associated with historical scholarship remain indispensable to assess in which cases mining big data for meaning is sound, productive and worthwhile from a humanities' point of view. Indeed, a word cloud is meaningless without the historian's ability to contextualise and understand the past 'from within'; and it takes a humanities scholar to understand whether a correlation that is statistically significant is also culturally relevant and historically meaningful. Moreover, no historian worth his salt will rely exclusively on text mining techniques, at least not anywhere in the near future. Text mining techniques will displace but they will not replace traditional hermeneutic methods. Indeed, Translantis explicitly includes conventional in-depth analyses as a way to explain the specific patterns, or parts of patterns, generated through computational methods: and of course, the source criticism that is part and parcel of the historical profession will remain a *sine qua non*, even if specific skills need to be honed to properly gauge the quality of a digital source. Historians do not have to worry that big data will ever replace intelligent inquiry, or that digital methodologies will serve as an alternative to historical theories. In fact, digital methods are worthless

without meaningful research questions and conceptual frameworks. One of the most important realisations is that even 'big' digitised data collections represent only a tiny and biased part of the historical evidence that historians have at their disposal, and using them requires constant awareness of their inherent biases and distortions. Among other things this is exemplified by the rather frustrating experiences with poor OCR quality and by the numerous lexicological challenges that still have to be met. We agree therefore, that the proper use of 'big data' requires equivalent quantities of critical sense and that we should stay clear from the dangerous illusion, as Andrew Prescott recently warned, 'that data can somehow be cut free from its historical moorings to enjoy an autonomous existence'.[35]

We might live in the age of the exabyte if not the zettabyte but it is important to acknowledge that only a sliver of our vast historical past is available in the form of bits and bytes. On a global scale, the current state of big data repositories hardly allows us to go beyond a still very circumscribed number of newspapers, a smattering of magazines and a painfully limited amount of digitised material that is as disparate as it is fragmentary. Furthermore, it remains to be seen whether different repositories can be accessed in a way that is useful to impatient scholars with limited time on their hands. Obstacles are not just technological in nature and do not only concern standards and protocols; excruciatingly complicated copyright issues are involved, as well as urgent concerns about OCR quality, textual stability, discrepancies within and between public and commercial providers of source material and the real threat that profitability may triumph over free access. At this moment we are not anywhere near to 'fingerprinting' different cultures in terms of attitudes, opinions and values – an ambition that is likely to revolutionise global history once it is realised. On the other hand, if these difficulties are surmounted – and there is no reason to believe that they cannot be – the opportunities for and possibilities of innovative historical research will be manifold.

There are things digital humanities research cannot do. One of them is to produce a historical narrative authored by a craftsman whose evocation of the past depends on individual erudition, scholarship, insight, talent and the ability to tell a story.[36] However, it is clear to us that the new tools and methodologies discussed in this paper, and to which all historians will soon have access, will be an important contribution to future historical scholarship. ◀

35 See for a recent assessment of these limitations, Andrew Prescott, 'Digital Riffs: The Deceptions of Data' (13 January 2013), http://digitalriffs.blogspot.co.uk/2013/01/the-deceptions-of-data.html.

36 Even here change is imminent, see Ann Rigney, 'When the Monograph is no Longer the Medium: Historical Narrative in the Online Age', *History and Theory* 49 (2010) 100-117.

Joris van Eijnatten (1964) is Professor of Cultural History at Utrecht University. He has worked on three overlapping and interrelated fields – the history of ideas, the history of religion, and the history of media and communication. He has written or edited more than ten books and authored around a hundred articles and contributions on subjects ranging from history of medicine and press freedom to the history of cultural values and public administration. Recent publications: co-authored with Fred van Lieburg, *Nederlandse religiegeschiedens* (second edition; Hilversum 2006) and *Hogere sferen. De ideeënwereld van Willem Bilderdijk, 1756-1831* (Hilversum 1998). He has a Dutch-language textbook *From Village Square to Cyberspace: A History of Communication, Media and Information* forthcoming (Prometheus). Email: j.vaneijnatten@uu.nl.

Toine Pieters (1960) is Professor of the History of Pharmacy at Utrecht University and senior fellow at the Descartes Centre for the History and Philosophy of the Sciences and the Humanities at Utrecht University. He has published extensively on the history of the production, distribution and consumption of drugs and the interwovenness with economy, science and the public sphere. His book *Interferon: The Science and Selling of a Miracle Drug* (London, New York 2005) explored the interaction of the broad range of actors-scientists, doctors, patients, journalists, government officers and executives of the pharmaceutical industry. He is research coordinator of multiple projects in the field of e-humanities. Email: t.pieters@uu.nl.

Jaap Verheul (1958) is associate Professor of Cultural History and Director of the American Studies program at Utrecht University. He has published on American and Dutch cultural history and on business history. His current research interest is American perceptions of Europe. He currently is finishing an intellectual biography of the American historian John Lothrop Motley (1814-1877). All three authors are applicants and coordinators of the digital humanities research projects Translantis: Digital Humanities Approaches to Reference Cultures: The Emergence of the United States in Public Discourse in the Netherlands, 1890-1990 (NWO-funded), and Asymmetrical Encounters: Digital Humanities Approaches to Reference Cultures in Europe, 1815-1992 (HERA-funded). Email: j.verheul@uu.nl.