2. **Shillingsburg, P. L.** (1991). *Text as matter, concept, and action*. Studies in Bibliography, 44:31–83.
3. **Shillingsburg, P.** (2006). *From Gutenberg to Google: Electronic Representations of Literary Texts*. Cambridge: Cambridge University Press.
4. **Caton, P.** (2013). *On the term text in digital humanities*. Literary and Linguistic Computing, 28(2):209–220.
5. **Gabler, H. W.** (2012). *Beyond author-centricity in scholarly editing*. Journal of Early Modern Studies, 1:15–35.
6. **Eggert, P.** (2009). *Securing the Past: Conservation in Art, Architecture and Literature*. Cambridge: Cambridge University Press.
7. **Tanselle, G. Thomas**. *A Rationale of Textual Criticism* Philadelphia: University of Pennsylvania Press, 1989. 104 pp.
8. **Sperberg-McQueen, C. M.** (2009). *How to teach your edition how to swim*. Literary and Linguistic Computing, 24(1):27–52.
9. **Robinson, P. M.** (2009). *What text really is not, and why editors have to learn to swim*. Literary and Linguistic Computing, 24(1):41–52.
10. **Robinson, P. M. W.** (2013). *Towards a theory of digital editions*. Variants 10. 105-132.
11. **Barthes, R.** (1968). *La Mort De l'Auteur*. Manteia (4e trimestre).
12. **Ong, J. Walter** (1975). *The Writer's Audience is Always a Fiction*. PMLA, Vol. 90/1, pp. 9-21
13. **McCarty, W.** (2005). *Humanities Computing*. Palgrave Macmillan.
14. **McCarty, W.** (2004). *Modeling: a Study in Words and Meanings*. in Companion to the DigitalHumanities. Blackwell.
15. **DeRose, S. J., D. G. Durand, E. Mylonas, and A. H. Renear** (1990). *What is Text, Really?* Journal of Computing in Higher Education 2:1 3-26.
16. **Huitfeldt, C.** (1994). *Multi-Dimensional Texts in a One-Dimensional Medium*. Computers and the Humanities, 28(4-5). Humanities Computing in Norway. 235-241.
17. **Pichler, A.** (1995). *Transcriptions, texts and interpretation*. In Johannessen, K. and Nor- denstam, T., editors, Culture and Value: Philosophy and the Cultural Sciences, pages 690–695. Austrian Ludwig Wittgenstein Society, Wien.
18. **Renear, A. H., Mylonas, E., and Durand, D.** (1996). *Refining our notion of what text really is: The problem of overlapping hierarchies*. In Ide, N. and Hockey, S., editors, Research in Humanities Computing. Oxford University Press.
19. **Pierazzo, E. and Stokes, P. A.** (2010). *Putting the text back into context: a codicological approach to manuscript transcription*. In Fischer, F., Fritze, C., and Voelgler, G., editors, Kodikologie und Palographie im Digitalen Zeitalter 2 - Codicology and palaeography in the digital age 2, pages 397–430. Books on Demand, Norderstedt.
20. **Deegan, M. and Sutherland, K.** (2009). *Transferred Illusions*. Digital Technology and the Forms of Print. Ashgate, Farnham.
21. **Sperberg-McQueen, C. M.** (2009). *How to teach your edition how to swim*. Literary and Linguistic Computing, 24(1):27–52.
22. **Huifeldt, C., and C. M. Sperberg-McQueen** (2008). *What is transcription? Literary and Linguistic Computing*. 23 (3). 295-310. doi:10.1093/llc/fqn013
23. **Huitfeldt, Claus, Yves Marcoux and C. M. Sperberg-McQueen** (2010). *Extension of the type/token distinction to document structure*. In Balisage: The Markup Conference 2010. held August 3-6, 2010 in Montréal, Canada. In Proceedings of Balisage: The Markup C
24. **Sperberg-McQueen, C. M.. Claus Huitfeldt, and Yves Marcoux** (2009). *What is transcription? Part 2*. Talk given at Digital Humanities 2009, College Park, Maryland. Slides on the Web at blackmesatech.com/2009/06/dh2009/.
25. **Caton, Paul** (2013=. *Pure transcriptional encoding*. Paper given at Digital Humanities 2013, Lincoln, Nebraska.
26. **Caton, P.** (2013). *On the term text in digital humanities*. Literary and Linguistic Computing, 28(2):209–220.
27. **Huitfeldt, Claus, and C. M. Sperberg-McQueen.** (2008) *What is transcription?* Literary & Linguistic Computing 23.3: 295-310.

Cultural text mining: using text mining to map the emergence of transnational reference cultures in public media repositories

Pieters, Toine
t.pieters@uu.nl
Utrecht University, The Netherlands

Verheul, Jaap
j.verheul@uu.nl
Utrecht University, The Netherlands

Introduction

This paper discusses the research project Translantis, which uses innovative technologies for cultural text mining to analyze large repositories of digitized public media, such as newspapers and journals.[1] The Translantis research team uses and develops the text mining tool Texcavator, which is based on the scalable open source text analysis service xTAS (developed by the Intelligent Systems Lab Amsterdam). The text analysis service xTAS has been used successfully in computational humanities projects such as Political Mashup, WAHSP, BILAND, and DutchSemCor. Within the context of the Translantis project, xTAS, coupled to Elasticsearch, will be further developed. Future versions will include clustering concepts and sentiment mining of issues in public debates. Translantis researchers are using Texcavator to detect and track cultural references in large textual corpora.

Use case: mining transnational references in public discourse

In order to test the potential of cultural text mining, Texcavator will be used to analyze the role of reference cultures in debates about social issues and collective identities. The central use case of this project is the emergence of the United States in public discourse in the Netherlands from the end of the nineteenth century to the end of the Cold War. This concept of reference culture is be used to discuss long-term asymmetrical processes of cultural exchange involving dimensions of power and hegemony. The concept recognizes the fact that some cultures assume a dominant role in the international circulation of knowledge and practices, offering or imposing a model that others imitate, adapt, or resist.

Reference cultures are mental constructs that do not necessarily represent a geopolitical reality with an internal hierarchy and recognizable borders. These culturally conditioned images of trans-national models are typically established and negotiated in public discourses over a long period of time. However, the specific historical dynamics of reference cultures have never been systematically analyzed and hence are not fully understood. To explore these dynamics, this project asks three interrelated questions.

1. How can e-tools be used to map trends and changes in relation to the economic power, cultural acceptance, and scientific and technological impact of the United States as reference culture?
2. How does public discourse reflect and influence the emergence and impact of reference cultures?
3. How were ideas, products and practices associated with the United States valued in Dutch public discourse between 1890 and 1990?

We propose that the key to understanding the emergence and dominance of reference cultures is to chart the public discourse in which these collective frames of reference are established. Text mining methodologies allow us to trace changes in “big data” repositories of public media, such as newspapers, journals, and other periodicals. Central to this project is the large digital data collection of the National Library

of the Netherlands (KB), which contains 9 million newspaper pages and over 1.5 million journal pages[2]. This large collection of serialized historical texts, which have been OCR-ed and provided with meta-tags, allows us for the first time to study long-term developments and transformations in national discourses in a systematic, longitudinal, and quantifiable way, by using innovative text-mining tools.

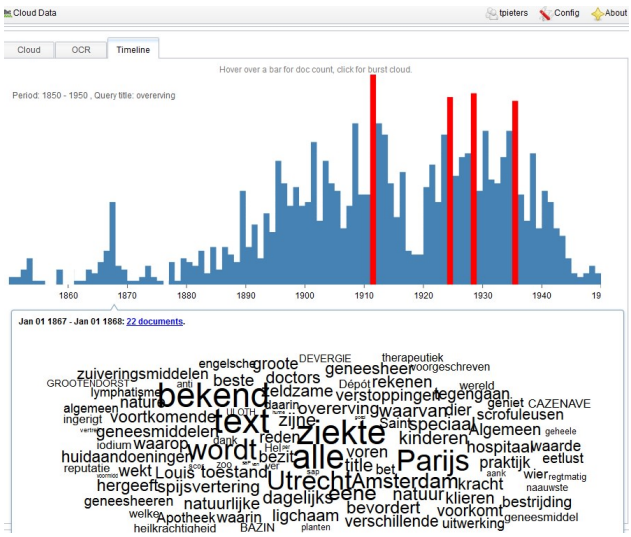
Methodological innovations and challenges

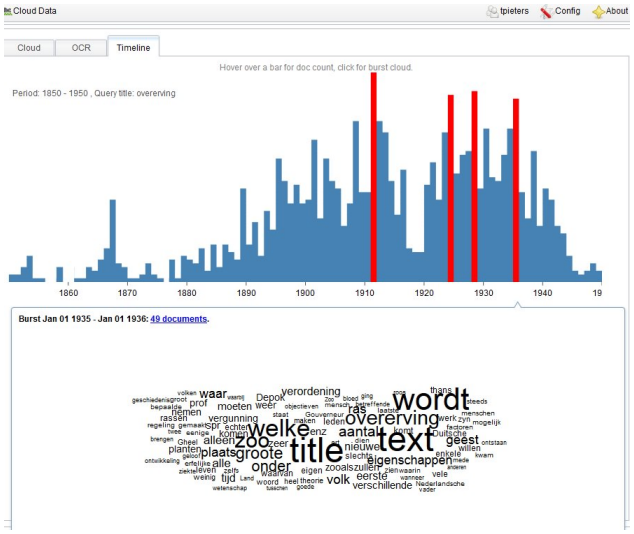
The semantic text mining tool Texcavator has direct access to historical textual repositories and is able to handle queries on-the-fly, and to produce visualization such as timelines and word clouds based on integrated topic modeling and NER modules. This allow us to test the value of qualitative heuristic models and to pair them in a meaningful fashion with quantitative methodology. Some of the methodological challenges involve the calibration between close and distant reading, the normalization of search results from unevenly distributed historical media, and adjusting for lexicological changes that affect the accuracy of sentiment mining and concept mining.

First results indicate the ability to mine “hidden debates” in public media in a bottom-up (inductive) manner, based on the footprints that used terms leave behind. More importantly, the tool is innovative in that it pinpoints continuities and discontinuities in public discourse, for instance by showing variations in the context in which key terms are used, and changes in sentiment values of words over time. We argue that this marks a promising transition from text mining to “concept mining” and new forms of cultural text mining that go beyond already established mining features.

Conclusions

We will demonstrate that semantic mining of big data open new vistas in historical research because they (a) provide a robust framework for producing new vistas on macro history; and (b) can be complemented with numerical data sets provided by other researchers, for example on economic and social trends. This, ultimately, is the transformative promise of digital humanities as a multi-dimensional window on political, economic, and social change.





References

1. *Translantis: Digital Humanities Approaches to Reference Cultures*; The Emergence of the United States in Public Discourse in the Netherlands, 1890-1990 (funded by the Netherlands Organization for Scientific Research), wwww.translantis.nl.

2. kranten.delpher.nl/

Aiden, Erez, and Jean-Baptiste Michel (2013). *Uncharted: Big Data as Lens on Human Culture*. New Yrok: Penguin.

Balog, K., M. Bron and M. de Rijke (2011). *‘Query Modeling for Entity Search Based on Terms, Categories and Examples,’* ACM Transactions on Information Systems 29, no. 4, Article 22, November 2011.

Dougherty, M., E.T.Meyer, C. Madsen, C. van den Heuvel, A. Thomas, and S. Wyatt (2010), *Researcher Engagement with Web Archives: State of the Art*. London: JISC.

Eijnatten, Joris van, Toine Pieters, and Jaap Verheul (2013). *“Big Data for Global History: The Transformative Promise of Digital Humanities.”* Low Countries Historical Review/BMGN 128-4: 55-77.

Hernandez, J.F., A.K. Mantel-Teeuwisse, G.J.M.W. van Thiel, S.V. Belitser, J.A.M. Raaijmakers and T. Pieters (2011). *Publication trends in newspapers and scientific journals for SSRIs and suicidality: a systematic longitudinal study.* BMJ OPEN 6, no. 1.

Huurnink, B., L. Hollink, W. van den Heuvel and M. de Rijke. *‘Search Behavior of Media Professionals at an Audiovisual Archive: A Transaction Log Analysis.’* Journal of the American Society for Information Science and Technology 61, no. 6 (June 2010): 1180-1197.

Huijnen P., F. Laan, M. de Rijke, and T. Pieters. *A digital humanities approach in the history of science; eugenics revisited in hidden debates by means of semantic text mining*. Histoinformatics, Springer (forthcoming, 2013).

Jijkoun, V., M. de Rijke and W. Weerkamp. *‘Generating Focused Topic-specific Sentiment Lexicons,’* 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), July 2010.

Meij, E., M. Bron, L. Hollink, B. Huurnink and M. de Rijke. *“Mapping queries to the Linking Open Data cloud: A case study using DBpedia.”* Journal of Web Semantics 9, no. 4 (November 2011): 418-433.

Pieters, T., and S. Snelders. *“Standardizing psychotropic drugs and drug practices in the twentieth century: Paradox of order and disorder.”* (2011) Studies in the History and Philosophy of the Biological and Biomedical Sciences 42: 412-415.

Snelders, S., and T. Pieters. (2011) *“Speed in the Third Reich: Metamphetamine (Pervitin) Uses and a Drug History From Below.”* Social History of Medicine. First published online: February 19.

Thomas, A., E.T., Meyer, M. Dougherty, C. van den Heuvel, C. Madsen, and S. Wyatt (2010). *Researcher*

Engagement with Web Archives: Challenges and Opportunities for Investment. London: JISC

Verheul, Jaap (2010). *"Through Foreign Eyes."* In Discovering the Dutch: On Culture and Society of the Netherlands, edited by Emmeline Besamusca and Jaap Verheul, 267-77. Amsterdam: Amsterdam University Press.

Aplicación del análisis dinámico de redes científicas al estudio de la evolución de la investigación española relacionada con el descriptor "historia del arte" durante 1976-2012, según ISOC.

Pino-Díaz, José
Dpto. Historia Arte. Universidad de Málaga, Spain

Cruces-Rodríguez, Antonio
antonio.cruces@uma.es
Dpto. Historia Arte. Universidad de Málaga, Spain

Rodríguez-Ortega, Nuria
nro@uma.es
Dpto. Historia Arte. Universidad de Málaga, Spain

Bailón-Moreno, Rafael
Dpto. Ingeniería Química. Universidad de Granada. Spain.

1. Objetivo

El objetivo de la presente comunicación ha sido detectar las dinámicas de la investigación española relacionadas con la historia del arte en un rango cronológico que se extiende desde 1976 hasta 2012. Para ello, se ha utilizado como estrategia el análisis dinámico de redes científicas, lo que nos ha permitido extraer de uno de los corpus más representativos de la investigación histórico-artística en España (ISOC) un conocimiento implícito sobre la evolución y desarrollo de las estrategias temáticas y líneas de investigación relacionadas con la historia del arte como descriptor.

La elección de este rango cronológico no es arbitraria, pues la segunda mitad de la década de los setenta se ha conceptualizado tradicionalmente como un punto de inflexión en los estudios histórico-artísticos en España, al incorporarse metodologías y modelos interpretativos foráneos una vez iniciado el periodo aperturista tras el fin de la Dictadura de Franco, la llegada de todo un conjunto de traducciones de obras que desarrollaban modelos interpretativos distintos a los anteriormente establecidos, y la consolidación de los estudios de Historia del Arte en España como campo académico autónomo.

2. Marco teórico-metodológic

Una red científica está formada según (Latour, 1983) por el conjunto de actores de la red (investigadores, centros, revistas, temas de investigación, etc.) y por el conjunto de asociaciones establecidas entre ellos. Los actores independientemente de su naturaleza pueden definirse siempre mediante palabras. La red de palabras y de asociaciones establecidas entre ellas es manifestación del comportamiento de la Sociedad y de la estructura del Conocimiento. Los actores son, igualmente, entidades dinámicas que continuamente redefinen sus relaciones y, en consecuencia, la red socio-cognitiva que conforman. Con el devenir, los actores y las relaciones cambian y dan lugar a nuevas redes, y así se suceden unas a otras a lo largo del tiempo.

Latour define la Teoría Actor-Red, teoría sociológica sobre la generación de conocimiento científico, como "Sociología de las Relaciones" (Latour, 2005). La teoría de la traducción (entendida esta como conversión, transformación, variación o cambio) estudia los cambios que se producen en las relaciones entre los actores de la red. Estos cambios en las relaciones entre los actores producen su aparición, fortalecimiento, equilibrio, debilitamiento o desaparición. La aparición de nuevos actores se produce por emergencia o por convergencia; el fortalecimiento se produce por convergencia o por evolución incremental; el equilibrio, por evolución estable; el debilitamiento, por evolución decremental o por divergencia; y la desaparición, por bifurcación o por exitus. A consecuencia de todos esos cambios, las redes tecnocientíficas de conocimiento se encuentran en continuo cambio (Ruiz-Baños, 1999).

Las relaciones naturales de coocurrencia de palabras en los textos científicos y técnicos constituyen una red tecnocientífica que puede ser analizada y cartografiada. El análisis y la visualización de estas redes es posible por el desarrollo de sistemas expertos denominados sistemas de conocimiento.

3. Criterios y proceso de análisis

El corpus documental de análisis lo constituyen los artículos de ISOC, http://bddoc.csic.es:8080/isoc.html, obtenidos mediante la búsqueda "historia del arte" en los campos título, resumen y descriptores, en el periodo 1976-2012. Este corpus está formado por 873 registros.

Se ha realizado el análisis dinámico de la red científica mediante el sistema de conocimiento Techné Coword; sistema de conocimiento que tiene su antecedente en Copalred (Bailón-Moreno, 2003), que a su vez tiene como precursor Leximappe (Law, Bauin, Courtial, & Whittaker, 1988); (Law & Whittaker, 1992). Para realizar el análisis de palabras asociadas, se ha creado un único campo formado por los descriptores, autores y revista de cada registro. Se han fijado los siguientes parámetros de análisis: ocurrencia mínima, 2; coocurrencia mínima, 2; tamaño mínimo de la subred, 2; y tamaño máximo, 12. Los subperiodos de estudio han sido: 1976-1983; 1984-1989; 1990-1995; 1996-2001; 2002-2007; y, 2008-2013.

Se han empleado técnicas KDD (knowledge databases discovery) y de text mining (coword analysis o análisis de palabras asociadas) para obtener las subredes de investigación más importantes y relevantes de cada subperiodo. El diagrama dinámico de subredes permite el análisis y la visualización de la evolución de los resultados de la investigación.

4. Resultados e interpretaciones

La gráfica de producción acumulada de documentos por año permite distinguir dos periodos de "inicio-expansión-agotamiento" sucesivos: uno de 1976 hasta 1986, y otro de 1987 hasta la actualidad (ver Figura 1)

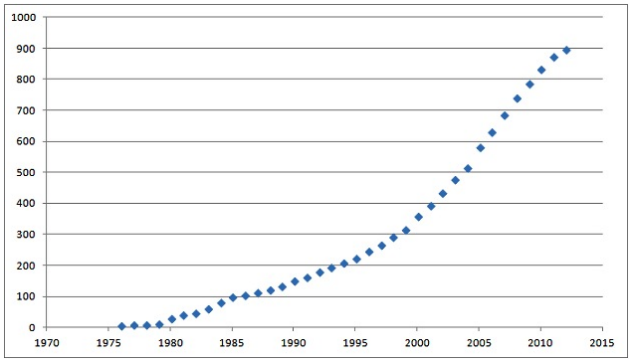


Fig. 1: Diagrama de producción anual acumulada de los documentos que incluyen "historia del arte" en su título, resumen o palabras clave.

Estos periodos son una constatación empírica de la Segunda Ley de Price o de Crecimiento Logístico de la Ciencia (de Solla Price, 1963), así como del concepto de cambio de paradigma, según la obra de Thomas Kuhn La Estructura de las Revoluciones Científicas (Kuhn, 1962). Esto último