art libraries journal 39 / 4 2014

Digitizing artist periodicals: new methodologies from the Digital Humanities for Analysing Artist Networks

Thomas Crombez

The research project Digital Archive of Belgian Neo-Avant-garde Periodicals (DABNAP) aims to digitize and analyse a large number of artists' periodicals from the period 1950–1990. The artistic renewal in Belgium since the 1950s, sustained by small groups of artists (such as G58 or De Nevelvlek), led to a first generation of post-war artist periodicals. Such titles proved decisive for the formation of the Belgian neo-avant-garde in literature and the visual arts. During the sixties and the seventies, happening and socially-engaged art took over and gave a new orientation to artist periodicals. In this article, I wish to highlight the challenges and difficulties of this project, for example, in dealing with non-standard formats, types of paper, typography, and non-paper inserts. A fully searchable archive of neo-avant-garde periodicals allows researchers to analyse in much more detail than before how influences from foreign literature and arts took root in the Belgian context.

Introduction

The Digital Archive of Belgian Neo-Avant-garde Periodicals is a digitization initiative by researchers at the Royal Academy of Fine Arts in Antwerp, Belgium. To digitize artist periodicals (and periodicals about art) is, in this case, not an aim in itself. We have a number of research goals we would like to achieve using the corpus. The most important objective is to examine the network of artists and artist groups that was behind the magazines, and how they were (or were not) connected to one another.

This implies the use of special, rather unorthodox methods of digitization. The project demands fast digitization, on a small budget, yielding a large-scale corpus of fully searchable text. At the end of my article, I will come back to these difficulties, but first I would like to present the corpus, using the tools we have at present, and then the theoretical issues and principles which have guided this project.

Presentation of DABNAP

The artistic renewal in Belgium since the 1950s, sustained by small groups of artists (such as G58 or De Nevelvlek), led to a first generation of post-war artist periodicals. Titles such as *Le surréalisme révolutionnaire*, *Cobra*, *De tafelronde*, *Het cahier* and *Gard sivik* proved influential for the formation of the Belgian neo-avant-garde in literature and the visual arts.

During the 1960s and the 1970s, happening and socially-engaged art (inspired particularly by the Provo movement) took over and gave a new orientation to artist periodicals. Examples include *Happening news*, *Revo*, *Anar*, *Milky way*, *Total's*,

art libraries journal

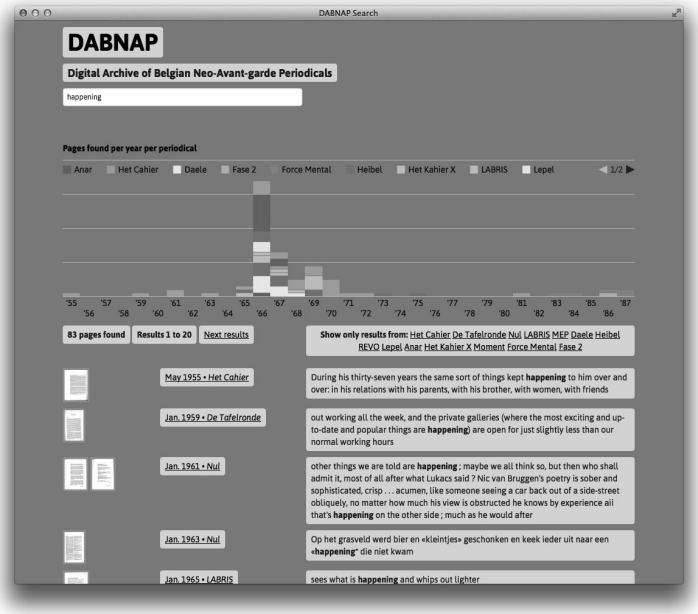

Fig 1. DABNAP Search Interface

and, on the side of literature, *Labris*, *Yang*, *Bok*, *MEP*, *Heibel*, *Boemerang*, and many others. Finally, the 1980s saw the rise of punk-inspired zines, including *Force mental*, *O* and *Fase 2*.

Most of these periodicals are currently present in the collection of the Hendrik Conscience Heritage Library (Erfgoedbibliotheek Hendrik Conscience, Antwerp) and the University of Antwerp Library, both partners in the DABNAP project. Where necessary, the library of the Museum of Contemporary Art (M HKA, Antwerp) and private collections are used to complement these collections.

The challenges and difficulties of this project lie in dealing with non-standard formats, types of paper, typography, and non-paper inserts. Paper sizes range from the ludicrously large (A2) to the very small (half of A5). Printing techniques include offset, mimeograph, screen-printing and photocopy, resulting in extremely diverse kinds of lettering and typography, which often confuses the OCR software that is used to extract text from the scanned pages.

The main questions I would like to address through the example of DABNAP are: how does digitization affect the value of periodicals for research into the history of art, literature and theatre? Ideally, what would researchers want their digital collections to be?

Digital Collections of Documents

Digital collections provide an excellent solution for the proverbial 'needle in a haystack' problems.

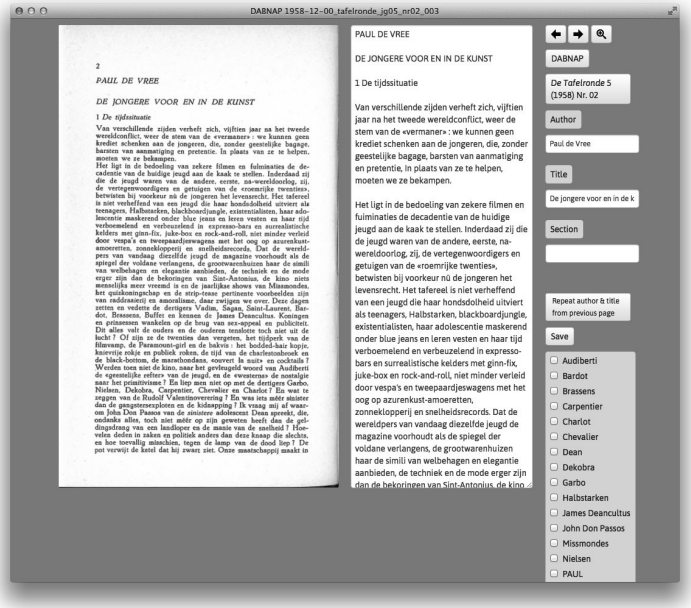

Fig 2. DABNAP Periodical Page View

A properly indexed digital collection makes the document sources not merely *browsable*, but fully *searchable*. Over the last few years, a number of noteworthy large-scale digitization projects of periodicals have been launched. One of the showcase projects of Google Books (the mass digitization project started in 2004) is the fully browsable archive of back issues from *LIFE magazine* (1936–1972). A second example is the Dutch Royal Library project *Historische kranten*, which was started in 2006 and has made available more than eight million pages from

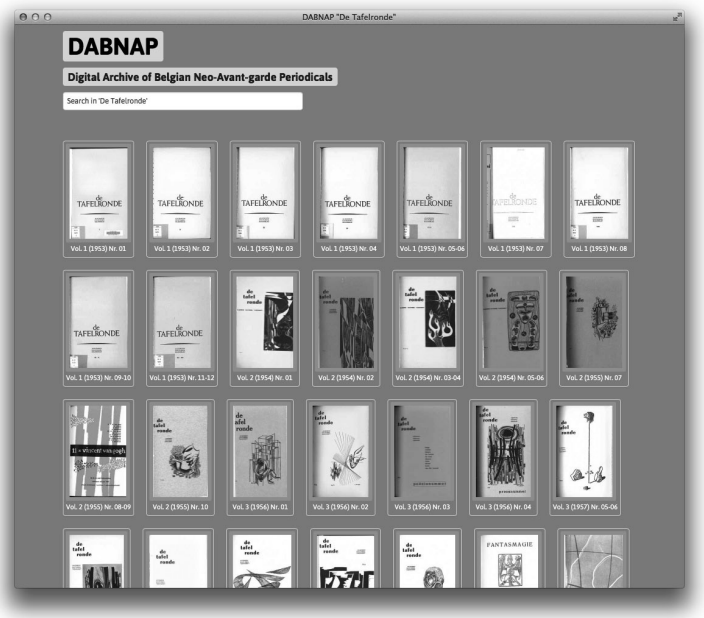

Fig 3. DABNAP Periodical Overview

Dutch-language newspapers of the seventeenth century up to 1995.

Such new resources will obviously have a huge impact on on-going and future historical research. Cultural history especially will benefit from the new availability of sources previously considered to be of second-rate importance because of their ephemeral nature. Still, these collections are conceived for a broad audience, and hence feature a general-purpose search interface. For instance, a search query in *Historische kranten* for articles from the inter-war period featuring the Amsterdam Stadsschouwburg (i.e., the city theatre) returns more than 31,000 results.[1] Every one of those results may be visualized in the context of the original newspaper page, but the historian cannot expect any extra help from the website in filtering the articles according to the theatrical events they announce, review, or advertise.

Similarly, Google uses basic text analysis to determine which words and expressions are characteristic for the book, or magazine issue, that the user is currently browsing. Take the February 1937 issue of *LIFE Magazine*. The interface is happy to inform me that 'Leon Trotsky', 'Reichstag', 'Studebaker', and 'Kleenex' are among the common terms and phrases. But it is obviously blind to the question whether these names belong to people, organizations, places, or brands. Neither does it provide a meaningful categorization of these expressions, but simply an unordered list.

Current digital archives, then, create at least as many problems as they solve. Language technology has become so pervasive in everyday life that users have come to foster very specific, but also quite restricted, expectations concerning searching and browsing. They prefer to direct their questions to a simple search box, as they have come to expect from the Google search interface.

The limitations of this model are evident in the recent efforts that Google itself has made to enrich search results with semantic metadata. On the current search results page, a number of options in the top bar or side bar of the screen allow one to narrow down the results according to a certain domain or type of results (for instance, news or images or books). Videos, photos and maps (if relevant) identify the search terms in a visual way. In 2012, the internet company introduced the Knowledge Graph, a wide-ranging database that is combined with thorough analysis of search queries in order to provide more 'semantic,' i.e., meaningful, information to users' queries. The system, for instance, now recognizes when a user is looking for a well-known name (for example, Leonardo da Vinci). It will try to provide not merely web pages in which that name is mentioned (*unstructured* text), but it will also give as much *structured* information as it has available. In case of da Vinci, the Google interface will present information on his biography (date of birth, place of birth, occupation, place of death), his family and his works.[2]

Semantic enrichment of data is also the approach commonly adopted by scholarly digitization projects. In contrast to mass digitization, specialized projects aim to provide more metadata, but they are necessarily much more limited in scope. Furthermore, in most cases such projects do not rely on automatic enrichment, but depend on editors who have manually marked up the digitized texts, adding semantic labels to certain documents, passages, or expressions. For example, the *Digitale Bibliotheek voor de Nederlandse Letteren*, or Digital Library of Dutch Literature (DBNL), collects more than 1,400 books and bound volumes of periodicals covering literature, history and linguistics written in Belgium and the Netherlands from the 13th century onwards.[3] Each digital document is encoded in TEI-XML and includes full bibliographical and editorial metadata. Such a digital corpus may be constructed manually when dealing with a limited set of documents. For large-scale projects, it would be unfeasible. But it is precisely these laborious parts of the work by researchers that stand in need of automation. Given the emergence of digital corpora of historical periodicals, could these sources be mined for information about artistic activity? Can we imagine and develop software that detects events in vast collections of text? In order to answer these ambitious questions, I would first like to take a closer look at the concepts of 'event' and 'document.'

From the Artistic Event to the Document as Event

Before a critical essay or a review becomes a document about an artistic practice, it is first a document about an artistic *event*. I will use the concept of the event as a springboard for re-thinking the nature of the periodical itself, in its capacity as a document. In other words, I would like to conceptualize *the document as event*, in order to transfer something of the dynamics of the event onto the document, which is commonly conceived of rather statically.

A periodical is occasioned by artistic events, such as the publication of new literary works, the exhibition of visual art, or the presentation of a new theatrical performance. However, as a document,

it can also be considered to be an 'event' in itself. The text functions as a linguistic meeting space for a wide diversity of named entities. This includes names of artists, writers, dramatists, performers, directors and critics; names of museums, galleries, theatres, companies and schools; and titles of books, art works and theatrical productions. In this sense, the totality of all artist periodicals and art-critical periodicals ever published (dispersed over countless of periodicals, many of which are probably already lost) could be said to hold a virtual record of modern art history. Even if only a small fraction of this textual treasure could be mined, it might produce a gargantuan database of past artistic events (e.g. theatre shows, art show openings, literary performances, musical performances, etc.).

Such an ambitious text-mining project is obviously difficult to realize using scholarly resources. The chief difficulty lies in having a sufficiently large set of digitized historical periodicals at one's disposal. Such collections are currently uncommon and, if available, not always very accessible. Still, given a large-scale digital corpus, it is possible and feasible to use off-the-shelf technology from the field of Natural Language Processing (NLP), more in particular information extraction, in order to (1) extract named entities from the digital text, and (2) detect meaningful relationships between those entities.[4] These relationships will often be specified in the document text, but they might also be implicit and presupposed. For instance, internet companies such as Google have developed data sets of relations about public figures based on pages from the English-language Wikipedia. Recently, Google released a data set of 50,000 such relationships, each of which has also been verified by five human assessors. For example, 10,000 of these instances detail a relationship of 'place of birth' between the name of a person and the name of a place.[5]

Text mining could thus be used as a tool for building a detailed database of artistic events, given a corpus of historical documents about art. This is not merely a chimerical vision; it is one of the ambitions of the DABNAP project. In order to elucidate the techniques that are going to be used, and to demonstrate which new questions may be answered through a semantically enriched corpus, I will discuss a case study from a previous digitization project.

A Case Study of Semantic Enrichment

Given a digitized collection of periodicals, which techniques can be used for automatically enriching their contents with extensive metadata? That such metadata might be of crucial importance to art scholars has become evident from the above. Semantically enriched data would allow searching for certain terms not merely in a strictly literal sense, but additionally according to the role they fulfil in the text. For example, it might become possible to query the Dutch newspaper collection *Historische kranten* for a certain theatre, say the Amsterdam Stadsschouwburg, and then ask the system to select all names from those articles that occur in the role of actor/actress. The result would consist of, approximately, the full group of performers that have played at the Stadsschouwburg in the period under review.

Two techniques from natural language processing and machine learning are crucial in the process of transforming plain-text documents into semantically marked up data. First, all named entities should be detected and extracted from the text, preferably according to their type (personal name, organization, place name). Second, the extracted names have to be classified according to their specific role. In the example below, I will focus on detecting *artistic roles of personal names* in the context of the performing arts (i.e. roles such as writer, artist, director, performer or critic). It is evident that the same methodology could also be used for distinguishing other names, such as organizational names (with roles such as: museums, galleries, art schools, theatre companies, playhouses, and funding bodies).

The digital corpus in question concerns the Flemish magazine *Etcetera*, which was started in 1983 in order to track innovative tendencies in Dutch and Flemish performing arts. The periodical has been instrumental in launching the international careers of artists such as Jan Fabre, Ivo Van Hove, Luk Perceval or Jan Lauwers. In 2011, the back issues of *Etcetera* for the period of 1983 to 2008 were fully digitized and made searchable by the Platform for Digital Humanities at the University of Antwerp, a project which was managed and developed by the author of this contribution.[6] The digital collection currently accounts for 114 issues, containing more than 2500 articles by 657 different authors. All pages were scanned, converted to text with OCR software, and manually corrected. In total, the corpus holds more than 5,000,000 words.

Named entities (people, places and organizations) were marked up automatically in the texts using Named Entity Recognition software for Dutch that was generously made available by the LT3 research group at the University of Ghent.[7] These additional metadata are presented to the user in the form of a side bar, which also contains the main metadata

(author, date, issue) and a selection of automatically generated keywords.

How can researchers use this extra layer of information in their research?

In the 1980s, a new and experimental generation of Flemish theatre directors was on the rise. This group included Jan Fabre, Jan Lauwers, Guy Cassiers, Ivo Van Hove, Jan Decorte and Luk Perceval. Together with the remarkable new élan in choreography (Anne Teresa De Keersmaeker, Wim Vandekeybus, Alain Platel), they became more widely known as the 'Flemish Wave'.

The term 'Flemish Wave' is a fascinating example of how canonization has operated in recent performing arts history. At first, the expression was used mostly in the Netherlands, because many of these young makers could find better working conditions there than in their own country. Later it was increasingly regarded as a mere marketing label, especially by local Flemish critics; but still the term has succeeded in establishing itself as an appropriate label for historicizing that episode.

It is a commonly held theory among Flemish theatre scholars, that *Etcetera* served to promote the Flemish Wave by giving much more attention to their productions than to the productions of their competitors in the state-funded repertory theatres (also known as city theatres) of Antwerp, Brussels and Ghent.

A semantically enriched corpus, such as the digital *Etcetera* archive, allows for testing such hypotheses. More particularly, I compared the mentions of this experimental group of theatre directors (Fabre, Lauwers, Cassiers, Van Hove, Decorte and Perceval) with a comparable group of mainstream directors working in the city theatres. How many *Etcetera* articles per year mention the names of the experimental group, and how many the names of the mainstream group?

The most conspicuous result of this analysis is the predominance of experimental theatre makers across the pages of *Etcetera*. The generation of Fabre and Perceval is mentioned in no less than five to ten per cent of articles per year. Jan Decorte and Jan Fabre especially stand out, being present in more than ten per cent of all articles over the entire period. During the 1980s, there are even some years when their mentions exceed 20%.

This stands in strong contrast to the lacunar presence of the mainstream artists associated with the much larger and amply subsidized city theatres. Seven out of nine of the mainstream theatre directors are rarely mentioned in *Etcetera* (i.e., in less than three per cent of articles on average). This is closely connected to the irregular distribution of the poorly scoring directors, which is the second significant trend. There seems to be a threshold between three and 6 per cent. Directors scoring less than 3% of mentions are also not present in every year, while directors scoring above 6% are more evenly distributed across the whole corpus.

Thus, the conclusion of the frequency analysis is the pronounced bias of the *Etcetera* editors and critics in favour of the experimental group of directors, most notably Jan Fabre and Jan Decorte. At some moment during the early 1980s, their respective interventions in the Flemish performing arts landscape must have seemed of such dramatic significance, that their names started to function as self-evident landmarks for the changes that were ahead. Merely by constantly featuring the experimental group of directors, and censoring more traditional figures, *Etcetera* has strongly promoted the Flemish Wave artists. (In the same way, *Theater Heute* may be said to have promoted the agenda of the *Regietheater*: not by publishing a *Regietheater* manifesto, but merely by reviewing certain productions, and not reviewing others.) Although there is no consciously articulated avant-garde programme in *Etcetera*, we can see that mentioning certain artists and leaving other ones out unconsciously articulates an avant-garde.

Conclusion

As the case study has shown, researchers can utilize collections of digitized periodicals to articulate and answer new historical questions. In order to determine how theatre criticism helped shape the new post-dramatic canon of directors, I analysed the digitized volumes of the Flemish periodical Etcetera to determine if either mainstream or experimental directors' names were mentioned most frequently. Turning back to the DABNAP project, I will conclude this contribution by summarizing the current state of the digitization process and how the project is expected to evolve with use. Currently, the project's ambition of digitizing circa 50 artist periodicals (comprising approximately 50,000 pages) has already been reached. Moreover, all of these have been converted into searchable text using OCR software. Only a small portion (circa 10,000 pages) of those pages has already been manually checked for remaining errors. A browser-based interface for consulting and querying the periodicals has already been developed. It also includes tools for visualizing the evolution over time of a search term.

One major handicap is that, since almost all textual and visual works from the periodicals are still under copyright, the digitized resources cannot be made available to the public right now, but are only accessible to a small group of researchers. The future of the project will almost certainly comprise the labour-intensive yet necessary process to identify the rights holders and ask for permission to make them available through a web interface.

References

1. Search query executed on kranten.kb.nl/ using Advanced Search (period: 1919 to 1939, search terms: 'Amsterdam Stadsschouwburg'). Accessed 16 June 2013.
2. Amit Singhal. "Introducing the Knowledge Graph: things, not strings" Online: *Official Google Blog*, 12 May 2012. Accessed 27 June 2013, http://googleblog.blogspot.be/2012/05/introducing-knowledge-graph-things-not.html.
3. See www.dbnl.org.
4. Jerry R. Hobbs and Ellen Rilott. "Information Extraction" in Nitin Indurkhya and Fred J. Damerau (eds.), *Handbook of Natural Language Processing*. New York: Chapman & Hall/CRC Press, 2010. Dan Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. New York: Prentice Hall, 2009.
   These two tasks are commonly designated as *named entity extraction* and *relation extraction*.
5. Dave Orr. "50,000 Lessons on How to Read: a relation extraction corpus." Online: *Google Research Blog*, 11 Apr. 2013. Accessed 15 June 2013, http://googleresearch.blogspot.be/2013/04/50000-lessons-on-how-to-read-relation.html.
6. See http://theater.uantwerpen.be/etc.
7. Ineke Schuurman, Véronique Hoste, and Paola Monachesi. "Interacting Semantic Layers of Annotation in SoNaR, a Reference Corpus of Contemporary Written Dutch". In Nicoletta Calzolari (Ed.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), 2010.
8. Geoff Nunberg. "Google Books: A Metadata Train Wreck". Online: *Language Log*, 29 Aug. 2009. Accessed 3 July 2013, http://languagelog.ldc.upenn.edu/nll/?p=1701.

Thomas Crombez
Koninklijke Academie voor Schone Kunsten
Mutsaardstraat 31
B-2000 Antwerp
Belgium
Email: thomas.crombez@ap.be