Using 'new' data sources for 'old' newspaper research: Developing guidelines for data collection[1]

PYTRIK SCHAFRAAD, FRED WESTER and PEER SCHEEPERS

Abstract

This article discusses the benefits and limitations of collecting electronic data for large-scale thematic content analysis. We will discuss a number of methodological and technical issues. The first one is the construction of a list of relevant keywords that serves as the primary data collecting device. This is not only a technical necessity, but also secures a theoretically and empirically valid collection of data. The second concern is the quality of electronic archive information. Finally, source-specific data characteristics and coding difficulties are dealt with. In conclusion, seven guidelines for electronic data collecting are proposed.

Keywords: content analysis, data collection, newspapers, electronic data sources

Introduction

In recent years, digital media have not only become a new field of research for media researchers. With the continuous growth of online resources they also offer new sources for data collection for researchers of 'old' media. The use of these data sources for data collecting implies new challenges as well as new chances and possibilities in the successive phases of the research process, from data collection to forms of automatic content analysis (see Krippendorf, 2004: 120). Electronic data collection is only one application of new facilities that have become available in recent years. A simultaneous development is the complete digitalization of content analysis in methods of automatic content analysis (see Kleinnijenhuis and Atteveldt, 2006). This article will shed light on the consequences of the use of these data sources for data collection, based on experiences from a pilot study on the newspaper coverage of the far right in the Netherlands and Flanders, using an all relevant cases

Communications 31 (2006), 455–467
DOI 10.1515/COMMUN.2006.029

03412059/2006/031–0455

data collection design. Are the possibilities of electronic data sources as unlimited as their advocates state? For our study we used two electronic data sources: Lexis Nexis Academic to collect articles from Dutch and German newspapers and Mediargus to collect articles from Flemish newspapers[2]. The emphasis will lie on the particularities of electronic data collection. In the process we will draw attention to the process of data collection for large-scale thematic content analysis in general, for this seems to be an underexposed step in the research process.

The leading questions in this article are: What are the benefits and limits of electronic data collection for large scale thematic and systematic-quantitative content analysis? And what guidelines can be explicated in order to carry out electronic data collection in a reliable and valid manner?

To some extent, electronic data collection does not differ from manual data collection in content analysis research. In most cases, the research questions will delineate what the topic of the selected articles is, from which newspapers (or other media), and from what time period the articles are collected. The second and third criteria are mostly explicitly dealt with in research reports. The operationalization of the theme or topic of the research, however, often remains unclear. We understand the word 'operationalization' here as the phase in the research process bridging theory and concepts to their practical application, in the data collection, research instrument, and analysis. The data collection often remains unexplained, even though it is a crucial step in the research process, as it is not arbitrary, but a process of selection and therefore relevant in terms of external validity. If the selection criteria are not formulated in detail, the outcome of the selection process will vary, once repeated. In collecting data for a study on coverage about the neighboring country in newspapers in the Netherlands and Germany, for example, this occurred when two assistants were asked separately to collect all articles on one news story (a football match between Germany and the Netherlands at the World Championship in 1974). The first assistant collected 35 articles from the same newspapers from which the second assistant only collected 17 articles (Wester, Pleijter, and Renckstorf, 2004; Pleijter, Renckstorf, and Wester, 2006). In the case of electronic data collection, this is an issue that also demands explicit operationalization for practical reasons.

In the following we will first elaborate on the operationalization of the main topic of a content analysis study, specified for electronic data collection. Next, we will discuss the benefits and limitations of electronic data collection. In the concluding paragraph, we will present seven guidelines/rules for electronic data gathering.

Demarcating the corpus: From theory and empirical manifestations to keywords

The resources for the operationalization of the main research topic into a data collection instrument are theoretical concepts and the manifestations of the topic in the public debate, all part of what Krippendorff calls the 'framework' (Krippendorff, 2004: 29−30). In our project, the research question concerns the variation in newspaper coverage of the far right in three countries (the Netherlands, Germany, and Flanders (Belgium)) between 1986 and 2004. Theories about the far-right party family and empirical information about the public debate on the far right are therefore important input for the operationalization, which will consist of a number of indicators that refer to the original research topic. From the theoretical concepts and the representations in the public debate, one needs to construct a list of keywords that includes all relevant aspects of the theoretical concept(s) as well as the variation in its manifestations in the public debate. In our research on the coverage of the far right, for example, synonyms of 'far right' (extreme right, national socialism), the most important characteristics of far-right ideology (racism, strong state, etc.), as well as relevant party and politician names had to be included in the keyword list (Kitchelt, 1997; Ignazi, 2003). The goal was to develop a keyword list that produces an exhaustive and relevant corpus, for we set out to collect all relevant cases.

The necessity of a theoretically and empirically informed keyword list

The obvious reason for applying the keyword list is its practical necessity when using an electronic data source, because one needs to type in keywords in order to select newspaper articles from it. Kleinnijenhuis and Van Atteveldt (2006: 229) emphasize that the research themes, operationalized in a keyword list, must be much more specified for electronic data collection than it would normally be.

However, we would like to elaborate on that emphasis. A theoretically and empirically informed keyword list is an important instrument for the construction of a valid, reliable, and reproducible corpus in all content analysis research on specific themes. It is an important step towards a more systematic approach of data collection for thematic content analysis. Data collection is not, and should not, be arbitrary. Similar to sampling in survey research, the procedures through which the corpus is constructed should be clear and reproducible for independent reviewers. Moreover, these procedures may provide empirical evidence for the population validity of the corpus. It is important to be explicit about the construction of the corpus in order to clarify the systematics

of the study (Wester, 2005: 10). In many research reports, this research phase is dealt with in half a sentence, which at the most mentions whether the material was gathered from archives, libraries, or electronic sources, wheras the size and contents of the corpus (and thus the results) actually depend on the search terms, keywords, or other collection criteria that have been used.

With the research goal and question(s) in mind, and the accompanying theoretical framework and information on its empirical manifestations at hand, one can assemble a procedure and set of criteria that enable the researcher to construct a relevant corpus in a transparent way. A theoretically and empirically informed keyword list may function as such a clear and reproducible set of selection criteria; if one or more of the keywords is/are present in the newspaper article, the article will be included in the corpus. Because the keyword list is directly based on key elements from the theoretical framework and their everyday manifestations, it also guarantees the internal and face validity of the corpus. Developing such a keyword list requires a strict procedure, one that involves testing the keyword list to exclude errors, or the possibility that the keyword list does not fit with the material. This last issue is especially important in research that involves longer time periods or different cultural contexts, where similar issues may be covered with different words[3]. Such a data collection method is similar to 'relevance sampling' (Krippendorf, 2004: 118), or 'theoretical sampling' (Alasuutari, 1995: 155). In practice the keyword list will not only be based on the theoretical framework, but also on empirical manifestations. The best reference for this type of data collecting might be 'purposive sampling' (Riffe, Lacy, and Fico, 1998: 86). The application of a theoretically and empirically informed keyword list (or an equivalent instrument for the construction of the corpus for content analysis around a specific theme) will clarify an often blurry phase in the research process and increase the validity, reliability, and therefore reproducibility of the study.

Sensitizing the keyword list

Once a keyword list has been constructed it should be sensitized by testing it for its productivity and selectivity. The way to do this is by applying it in an electronic data source. The keyword list will then produce a certain output. One then needs to analyze which keywords derive which kinds of articles from the electronic data source. The original keyword list can be divided into three categories of keywords:

1. Productive and selective keywords (in our example: Vlaams Blok, racism);

2. Productive but unselective keywords (in our example: asylum seeker, criminal justice);
3. Unproductive keywords (in our example: taboo, racist violence).

The first group of keywords should be maintained. The second group should be replaced with possibly more selective synonyms, or, if these are not available, be excluded from the keyword list. In our case, the keyword 'asylum seeker' was very productive, and most of the articles it produced concern asylum seekers, but not the far-right or even any political standpoint in the asylum debate whatsoever. Thus, it had to be removed; all unproductive keywords must be deleted from the keyword list.

The new list has to be tested again through the same procedure. This may have to be repeated various times. A good final test is to re-search the output of the keyword list in the paper edition of the same newspaper to see if the keyword list produces all relevant articles from the entire paper (exhaustiveness and relevance of the corpus). In our example, 151 articles found with the keyword list in LNA were traced back manually in the paper version of *de Volkskrant*. All pages of the paper version were scanned for additional articles on the far right. This test did not reveal any blind spots in the keyword list. In the test case, the keyword list turned out to be exhaustive. No additional relevant articles were found in the thorough search in the paper versions. However (this is an additional issue), the level of effectiveness of the keyword list cannot be brought up to 100%, that is, there is always overselection. The corpus that it produces will always have to be manually checked for irrelevant articles (e. g., published lists of MP candidates, or duplicate articles when an article is found twice in the data source)[4]. A keyword list that has been constructed and sensitized through this procedure can be found in Appendix I.

Beneficial and problematic issues of electronic data collection

Once a selection instrument (keyword list) is available, the actual selection of articles can take place. Applying electronic data sources such as Lexis Nexis Academic (LNA) or Mediargus (Med) is then beneficiary for a number of reasons.

First of all, with a valid selection instrument it is an exhaustive method of collecting all relevant newspaper articles. The computer does not 'forget' or 'overlook' even the smallest article. The most obvious advantage is the time and labour saving that can be reached with the application of electronic data sources. This is true, although not as much as it at first seems, because the necessary second order manual selection still

takes a lot of time. Often the selection of articles takes almost as much time as the coding process later on (Hijmans, Wester, and Pleijter, 2003). The workload should still not be underestimated, but fact is that complicated time-consuming acts with newspaper books on photocopiers in libraries are unnecessary. Instead, downloading the selection from an electronic data source directly results in a digital database of newspaper articles, which is a preferable storage method that also opens possibilities for automatic content analysis and easier sharing of datasets.

Two other beneficiary points do not directly concern the data collection itself. First of all, sensitizing the keyword list, as described above, would be a much more labor-intensive task if done manually. Using electronic data sources to carry out this task is much faster and more efficient. And even if one is not exactly sure what would be the most suitable period of time to focus on, this can easily be tested by carrying out a search with a number of keywords or keyword combinations in different time frames, in order to find out in which time frame a certain issue is most relevant. Van Praag (2005: 31) also mentions this use of electronic data sources: Frequencies analysis (frequency of appearance of keyword(s) combinations) has a strong selection value for newspaper articles that are worth detailed analysis.

Limitations

Despite these benefits, the application of electronic data sources for data collection is not entirely without problems. There are some limitations that are rooted in the fact that these data sources have not been designed specifically for content analysis usage. There are two main issues.

Firstly, electronically selected articles are not the same as their paper versions. That is why there are limitations to the questions that can be answered with electronically selected newspaper articles. The newspaper articles are presented in either ascii or xml format. The electronic data sources contain all individual article texts as published in the original newspaper, but other information is lost. All visual parts of the newspaper, such as images, figures, and tables are not available. Moreover, other visual and context information is also lost. There is no information on the exact place of the article on a page, no information on its 'news context', the size of headlines is unknown, and formal characteristics (page number, section title, author name) are not always complete and reliable (see below). This all means that some questions that content analyzers (d'Haenens and de Lange, 2001; Hijmans, Pleijter, and Wester, 2003; Van Gorp, 2005, 2006; Wester, Pleijter, and Hijmans, 2006) tend to ask cannot be answered based on an electronically gathered corpus (variables such as the size of the headline and the content of accompany-

ing images cannot be questioned). Another consequence of the lack of sufficient formal information is the appearance of coding difficulties with these last issues. In the case of manual selection from paper versions of the newspaper, it is obvious when something is a 'letter to the editor', because it is directly taken out of that section in a specific layout. In the case of electronic data-gathering the genre may not be as easily recognizable as such; 'it looks different.' Coders therefore need additional instructions on how to recognize various genres in electronically collected newspaper articles.

The second problem is that electronic data gathering does not mean manual selection is completely banned. The selection of newspaper articles from electronic data sources will always have to go through two stages: an electronic first (rough) selection and a secondary manual (fine) selection. This is unavoidable due to two aspects of electronic data collection. The delineation of a corpus is often more specific than the possibilities of the LNA/Mediargus search machines allow. Out of all articles that include one or more of our keywords, we only need those from the news sections, opinion and debate section, and election specials, and not, for example, book reviews. In other examples, researchers are only interested in front page news or the editorials. These further limitations cannot be entered/defined in the search machine of the electronic data sources and therefore need to be carried out manually after the initial electronic search with the keyword list. Neuendorf (2002: 92, 223) has also mentioned this limitation, 'warning' researchers that searches are still not fully automatic and are very much a creative process. In the next section and Appendix II, we will open the black box of that creative process. The other thing is that the keyword list, no matter how fine-tuned it is, will always produce irrelevant articles due to the wide (and sometimes unexpected) use of many of the keywords (in our own example, the name of party leader Pim Fortuyn produces a lot of irrelevant articles containing the phrase "… na de moord op Pim Fortuyn ..." [after the murder of Pim Fortuyn]). Therefore, a careful second selection is needed, which means the researcher has to formulate selection criteria and apply them to the rough selection obtained electronically[5].

Conclusion: Seven guidelines for electronic data collection

Electronic data collection is not beneficial in all cases. But if the research project meets three simple demands, it is a route well worth taking. It is exhaustive, time- and labor-saving, results in a exhaustive digital database, and enables easy sensitizing of the keyword list and easy determination of relevant research periods. The three requirements are as follows:

1. Since preparation is an all determining and labor-intensive business, benefits are more effective in the case of large-scale thematic content analysis.
2. The content analysis must concern the verbal content of (newspaper) articles, because visual content, formal and layout aspects, and context information of the articles are not included, or at least limited in the data sources.
3. The research should concern relatively recent affairs. Even though it varies strongly between newspaper titles, the data sources contain only newspaper volumes dating back to the mid-nineties (see also Neuendorf, 2002: 221)[6].

Drawing from our empirical experience, we conclude this investigation of the challenges of these data sources with six guidelines that enable the researcher to draw a corpus from electronic data sources in a systematic way that is well rooted in theoretical and empirical assumptions and reproducible for colleague scientists.

1. Make an overview of key concepts out of the theoretical and historical literature on the subject.
2. Extend this overview with the manifestations of these key concepts in the public debate, especially (but not exclusively) the media that will be studied.
3. Operationalize the overview into a selection instrument in the form of a keyword list. This keyword list enables systematic and purposive search through the data sources.
4. Test the keyword list on its productive and selective power in order to assure accuracy and avoid 'irrelevant production'. Especially the development of the keyword list is a process of many minor and major changes and try-outs. Also, the list of criteria as mentioned in Appendix II becomes clear during the process of data-collecting and therefore needs careful documentation.
5. Formulate specific guidelines for a manual (second order) selection based on knowledge of the raw data and research question. These need specification for each individual research project[7]. Explication of the development of the selection instrument and selection criteria enables the researcher to strengthen the arguments and conclusions in a research report (Wester, 2005: 13) and thereby increase accountability, reliability, and data-collecting validity (Krippendorf, 2004: 319). Therefore, it is important to document all steps in the process (Pleijter, 2006: 156−158).
6. Carry out the second order manual selection after application of the instrument in order to exclude articles from sections that one does

not want to include in the research, or articles that contain keywords outside the relevant context. These criteria should be seen as the completing part of the data-collection instrument, besides the keyword list. To avoid differences in outcomes of the application of the guideline for this selection between different people working on the data collection, the researcher should first organize a 'data-collector training', similar to a coder training, which is usual in content analysis research for the same reasons (see Neuendorf, 2002: 133). One could even consider an 'interdata-collector test' to further secure reliability.

7. Pay attention to the specific characteristics of electronically selected material in the coding instructions and coder training. Knowledge of these characteristics (as described in section 3) is needed to be able to code this material correctly.

In conclusion, these electronic data sources may not have been sent from heaven, but are very beneficiary for data collection once a few guidelines are applied.

Acknowledgement

We wish to thank the Netherlands Organization for Scientific Research (NWO) for their financial support.

Notes

1. An earlier version of this article was presented at the European Communication Conference in Amsterdam on November 25th, 2005.
2. See for background information on LNA and a users guide Neuendorf (2002: 220−221). See Lexis Nexis Academic (library account): http://www.lexisnexis.nl/lnn/ln/KUN/ and Mediargus (private account): http://www.mediargus.be.
3. Such is the case, for example, with the immigration issue. While in the seventies the Dutch word 'gastarbeider' (guestworker) would be appropriate to find most articles on the issue, in the nineties 'allochtoon' (migrant) would serve better and in the Flemish context 'allochtoon' is hardly used, instead 'migrant' is a much more common concept in the Flemish public sphere.
4. At various occasions we found the exact same article twice in one search result in LNA or in Mediargus. We have not tested how often this form of over selection occurs. It is our expectation that this will be less than 1%. However, it is difficult to verify this because in most instance it is only notified during the second order selection if the duplicates are located close to the first copy of the article in the search result list.
5. This second, manual step may become undoable in case of extremely large corpuses ($> n = 10.000$). In those cases it would be too much work and the researcher should then alter (often lower) the criteria for inclusion in the corpus.
6. A general guideline is that so-called quality papers date back to sometime in the early nineties, whereas most popular papers or tabloid date back no further then

1998. LNA and Mediargus contain the newspapers we used in our research from the following years on: *de Volkskrant* (1995), *NRC Handelsblad* (1991), *De Telegraaf* (1999), *De Standaard* (1995), *De Morgen* (1995), *Het Laatste Nieuws* (1998), *Frankfurter Allgemeine Zeitung* (1993), *Süddeutsche Zeitung* (1995), *Bild* (not at all).

7. See Appendix II for the general selection rules and our application of these criteria.

References

Alasuutari, P. (1995). *Researching culture. Qualitative method and cultural studies.* London: Sage.

D'Haenens, L. and Lange, M. de (2001). Framing of asylum seekers in Dutch regional newspapers. *Media, Culture and Society, 23*(6), 847−860.

Gorp, B. van (2005). Where is the frame? Victims and intruders in the Belgian press coverage of the asylum issue. *European Journal of Communication, 20*(4), 484−507.

Gorp, B. van (2006). *Framing asiel. Indringers en slachtoffers in de pers*. Leuven: Acco.

Hijmans, E., Wester, F., and Pleijter, A. (2003). Covering scientific research in Dutch newspapers. *Science Communication, 25*(2), 153−176.

Ignazi, P. (2003). *Extreme right parties in Western Europe.* Oxford: Oxford University Press.

Kitschelt, H. (1997). *The radical right in Western Europe. A comparative analysis.* Ann Arbor, MI: University of Michigan Press.

Kleinnijenhuis, J. and Atteveldt, W. van (2006). Geautomatiseerde inhoudsanalyse, met de berichtgeving over het EU-referendum, als voorbeeld. In F. Wester (Ed.), *Inhoudsanalyse: Theorie en praktijk* (pp. 227−250). Utrecht: Kluwer.

Krippendorff, K. (2004). *Content analysis. An introduction to its methodology.* London: Sage.

Neuendorf, K. (2002). *The content analysis guidebook.* London: Sage.

Pleijter, A. (2006). *Typen en logica van kwalitatieve inhoudsanalyse in de communicatiewetenschap.* Ubbergen: Tandem Felix.

Pleijter, A., Renckstorf, K., and Wester, F. (2006). Materiaalselectie en registratie: Berichten over het buurland in dagbladen uit de Duits-Nederlandse grensstreek. In F. Wester (Ed.), *Inhoudsanalyse: Theorie en praktijk* (pp. 45−63). Utrecht: Kluwer.

Praag, P. van (2005). De veranderende Nederlandse campagne cultuur. In K. Brants and P. van Praag (Eds), *Politiek en media in verwarring. De verkiezingscampagnes in het lange jaar 2002* (pp. 19−43). Amsterdam: Het Spinhuis.

Riffe, D., Lacy, S., and Fico, F. (1998). *Analyzing media messages. Using quanititative content analysis in research.* Mahwah, NJ: Lawrence Erlbaum.

Wester, F. (1995). Inhoudsanalyse als systematisch kwantificerende werkwijze. In F. Wester, K. Renckstorf, and H. Hutten, *Onderzoekstypen in de communicatiewetenschap* (pp. 134−162). Utrecht: Bohn, Stafleu and Van Lochem.

Wester, F. (2005). De methodenparagraaf in rapportages over kwalitatief onderzoek. In F. Wester, H. Boeije, and T. Hak, *Methodische keuzen in kwalitatief onderzoek. Tweede Kwalon Jaarboek* (pp. 8−14). Utrecht: Lemma.

Wester, F. (2006). Inhoudsanalyse als onderzoeksontwerp. In F. Wester (Ed.), *Inhoudsanalyse: Theorie en praktijk.* Utrecht: Kluwer.

Wester, F., Pleijter, A., and Renckstorf, K. (2004). Exploring newspaper portrayals: A logic for interpretative content analysis. *Communications, 29,* 495−513.

Wester, F., Pleijter, A., and Hijmans, E. (2006). Instrument en codeerformulier; Wetenschap in de krant. In F. Wester (Ed.), *Inhoudsanalyse: Theorie en praktijk*. Utrecht: Kluwer.

Appendix I: A keyword list

Below you will find the keyword list as it has been used and developed in the example research that is mentioned in this article. Two versions are shown. One is prepared for the Flemish newspaper articles in the election period of 1999; the other for newspaper articles from the period of the Dutch elections of 1998. Since the project involves researching Dutch, German, and Flemish newspapers in the election years between 1986−2004, the national part of the election list is specified for each national context and period (with specific names of far-right politicians, specific topics, and specific keywords for general topics (as, for example, migration involves 'guest workers' in one period and 'asylum seekers' in the other). The keyword list was made in three different versions, one for each country in the project. It consists of a standard list with for each country an additional short list of country-specific keywords (e. g., the Flemish list contains 'cordon sanitaire' and 'commonitaire' as extra keywords referring to specific Belgian issues). This was done because in each country the far right has a specific meaning and representation. The list of party, organization, and personal names of relevant far-right actors is also country-specific, as well as time-period-specific. The standard part of the keyword list consists of operationalizations of theoretical (and empirical) key concepts and their synonyms. The lists below are specifically prepared for direct use in LNA (General and Dutch keywords) and Mediargus (Belgian keywords).

Appendix Ia: The general keyword list
'Extreem rechts', extreemrechts, racis*, fascis* 'neo nazi', rechtsextreem, nationalis*, 'nationaal socialisme', xenofo*, vreemdelingenhaat, integratie, assimilatie, immigratie, apartheid, 'anti semitisme', 'sans papiers', populisme, 'gevestigde orde', proteststem, gezinspolitiek, holocaust, 'anti feminisme', 'anti communisme, partijverbod, tegendemonstratie.

Appendix Ib: The keyword list for the Dutch 1998 case
A. Names. Centrumdemocraten, CD, CP '86, NVU, Nederlands Blok, Voorpost, Antifa, KAFKA, 'Nederland Bekend Kleur', Janmaat, Freling, Glimmerveen, Wim Vreeswijk, Wim Beaux.

B. Issue keywords. Vol is vol, Kedichem.

Appendix Ic: The keyword list for the Belgium 1999 case
A. Names. 'Vlaams Blok', 'Charta 91', 'VAKA/Hand in Hand', 'Gerolf Annemans', 'Dewinter', 'Frank Vanhecke', 'Karel Dillen', 'Phillip de Man', 'Francis van den Eynde', 'Alexandra Colen', 'Johan Demol'.

B. Issue keywords. IJzerbedevaart', 'Vlaamse beweging', 'splitsing sociale zekerheid', 'eigen volk eerst', '70 punten programma', 'ethische partij', 'stemrecht voor migranten', 'cordon sanitaire'.

Appendix II: Protocol for electronic data selection

General	Applied
A. Developing a keyword list	
• Derive an overview of relevant characteristics of central themes of research from the theoretical framework. ⬇	
• Extend the overview with manifestations of these key concepts in random sample of newspaper articles and extra-medial material in order to guarantee theoretical and empirical relevance. Define key concepts and synonyms. ⬇	
• Operationalization in keyword list (first version). ⬇	
• Test productivity and selectivity of the keyword list by applying to one or more small samples. ⬇ ⬆	
• Second version. ⬇	
• Repeat testing until the list of keywords has reached a sufficient level of productivity and selectivity. ⬇	
• Carry out a search with the keyword list in an electronic data source and compare outcomes to a paper search. If the paper search does not reveal relevant articles that have not been found with the keyword list, the keyword list is ready for application. ⬇	
• Final version.	

B. Selection of the articles/sampling	
1. Carry out a search within the set time period and newspaper title with the complete keyword list. When searching paper or MF data, scan headlines, leads, and then paragraphs for keywords.	• Electronic: Type in the keywords in 'search term' box and define title and dates; • Paper/MF: Scan pages of the newspapers in the given time period for the keywords.
2. Save the resulting rough selection.	• Electronic: There is an option for downloading the selection; • Paper/MF: Print/copy the selected articles.
3. Scan the rough selection by hand, delete irrelevant articles based on explicated criteria: • the delineation of newspaper sections • topic relevance • possibly included disambiguaties • case specific criteria (e. g., related to critical events)	Exclude for example: • articles with the last name of a politician that appears to deal with another person with the same name; • lists of (chosen) MP candidates; • foreign news covering other countries than the Netherlands, Belgium, or Germany; • articles from the literature sections (book reviews); • articles from the culture and leisure sections: CD reviews (especially in Dutch papers); • reviews of plays, films, and other cultural events, sections not mentioned in step 2; • articles on the dioxine affair in the Flemish papers. • specific additional criteria on articles on Fortuyn and LPF.
4. Save final selection.	• Electronic: save each article starting on a new page (needs editing in the case of LNA) and print for filing; • Paper/MF: file.