Laboratories of Community: How Digital Humanities Can Further New European Integration History

Mariona Coll Ardanuy[1(✉)], Maarten van den Bos[2], and Caroline Sporleder[1]

[1] University of Trier, Universitätsring 15, 54296 Trier, Germany
{s2macoll,sporledc}@uni-trier.de
[2] University of Utrecht, Campusplein 1, 3584 ED Utrecht, The Netherlands
m.j.a.vandenbos@uu.nl

Abstract. It has been said that media is an important but mostly overlooked player in European integration history. Now, the mass digitisation of newspapers and the introduction of new digital techniques promise great potential to remedy this inattention. With the conjecture that people are drivers and carriers of change, we propose a people-centric approach to mine news articles in a way that can be most useful to further historical research. In this paper, we describe a methodology for building social networks from unstructured news stories, with the European integration scenario serving as a case study.

Keywords: Digital history · Text mining · European integration · International relations · Social network · Public discourse

1 Introduction

The historiography of European integration can be divided into three phases [21]. In the seventies, historians from a broad spectrum of backgrounds such as diplomatic history, economic history and the history of ideas began to analyse the origins of European cooperation. Subsequently, from the second half of the eighties, this fragmented field began not only to integrate but also to interact across disciplinary boundaries with international relations theory, judicial history and political philosophy. In the first half of the nineties, the international relations perspective gained dominance framing Europe as a political entity theoretically situated somewhere on a scale between a federation of states and a federal state. The third phase took off with a devastating critique of the dominance of the state as actor in the history and theory of international relations. Building upon the theoretical insights of early constructivists like Alexander Wendt, some scholars argued for not only a transnational history of the public opinion on the European project, but also for a shift of focus from interstate relations and government policy towards non-political actors, public discourses and popular images of Europe.

L.M. Aiello and D. McFarland (Eds.): SocInfo 2014 Workshops, LNCS 8852, pp. 284–293, 2015.
DOI: 10.1007/978-3-319-15168-7_36

The mass digitisation of books, newspapers and other historical materials and the introduction of new digital techniques promise new possibilities in furthering our knowledge of European integration. We propose a method to automatically extract networks of people mentioned in news stories, weighted according to their significance in the news and distributed according to their co-occurrence in the text. We suggest a simple strategy to incorporate shared contextual information for each pair of nodes in the network, based on word counts and tf-idf frequencies, implemented as edge attributes. The aim of the networks is to allow historians to have not only a bird's eye view of the scenario of a certain period of time (in our case, 1945–1955), but also to have an overview of which relevant topics were discussed when a certain historical actor was mentioned. By reducing the whole corpus of news into a network structure, we find some expected results and, more interestingly, some unexpected results. In this regard, we think of our method as a generator of hypotheses.

In this particular case study, we look at articles on the first steps of the integration process in the late forties and early fifties in the Netherlands. As one of the founding members of the European Union, the Netherlands proves an interesting case because of its diverse media landscape, in which different socio-political communities had their own ideological media outlet that coexisted with more neutral, general newspapers [2]. In a recent article, Hans-Jörg Trenz has named the media as an important but mostly overlooked player in integration history [22]. By using digitised newspapers from the large repository of the Dutch Royal Library to extract social networks and their main topics of conversation, we will develop a method to further the history of European integration in a digital fashion.

2 State of the Art

2.1 Historiographical Background

Theory and historiography of the European Union until the late nineties can be roughly divided into two fields of interest. In the first, the focus is on the current European institutions and treaties. In the second, the history of European integration is seen through the lens of international relations theory [26]. A landmark for this dominance was the publication of Alan Milward's *The European rescue of the Nation State*, in which he developed the argument that 'integration was not the supersession of the nation-state by another form of governance as the nation-state became incapable, but the creation of the European nation-states themselves for their own purpose, as an act of national will' [14]. From a more theoretical perspective, the leading scholar on European integration Andrew Moravcsik agreed as he concluded that the integration process did not supersede or circumvent the political will of national leaders, but reflected it [15].

Both their approaches however came under scrutiny in the late nineties. In his *Social Theory of International Politics*, Alexander Wendt developed a theory of the international system as a social construction. Following in his footsteps, a whole series of so-called constructivist studies focussed on the role of non-state actors, civil society and public opinion in international relations [28] [27].

Within the context of European integration historiography, constructivism got a firm boost from the 2005 rejection of the Treaty establishing a Constitution for Europe. Signed in October 2004 and quickly ratified by some member states, the ratifying process stranded only a summer later on the rejection of the treaty by French and Dutch voters. Especially the high turnout in both referenda sparked academic interest in the public image of and popular support for European integration [6] [16].

In recent years, a lot of work has been done broadening the research agenda by implementing transnational approaches, the history of interest groups, European bureaucracy and civil society organisations. In an attempt to set the future research agenda, the leading scholar Wolfram Kaiser has made a convincing plea for a network approach of the history of European integration. He introduced the concept 'epistemic communities' as a possibility to further what he labelled new European integration history. These communities essentially are networks of people who share expert knowledge and have a common understanding of a certain issue. Therefore, they function as 'channels through which new ideas circulate from societies to governments as well as from country to country' [27].

For historians, this network approach has distinct advantages over the other attempts to conceptualize the history of European Union. It is no rigid theory on the origins, functioning or development of European integration, but merely a heuristic concept that draws our attention to public debate, the emerging of new policy networks and the transfer of ideas and concepts across socio-political, academic and national borders. On the downside is the problem of operationalization. Networks of experts are only sometimes explicit (think tanks, for instance), work mostly behind the scenes and can be in competition with other networks; they are hard to find using traditional research strategies.

2.2 Computational Background

In the field of European Union studies, computational techniques to map out public discourse are, with counted exceptions [29], still in the earliest stage. A recent book on the role of national self-images in the perception of European integration in England, Germany and the Netherlands [5] uses a wide selection of more than a thousand editorials, but selection and analysis have been done merely by hand. Other studies that convincingly prove the investigatory value of newspapers also use traditional techniques to select and analyse the source material [12] [13]. The utility of using social networks in historical research has been demonstrated by several studies [18] [7] [20]. However, most of these studies create the networks either manually or drawn from structured data.

With the explosion of big data, there is a pressing need to improve techniques that address information extraction from unstructured data, which make for most of the real-world data with which historians have to deal. Text mining and other natural language processing techniques have been seeking since their beginnings a solution to the problem of unstructured data. In our method, we propose an entity-centric analysis of the data, an approach that has gained growing popularity in quantitative literary analysis [4] [17] [1] [3]. Some of these

studies represent novels as social networks of characters that typify the skeleton representation of the plot. The self-containing nature of the literary works makes for the biggest difference between fiction and real world data. When a novel ends, its characters cease to exist. When working with news data, we are not in a microcosmos anymore, and thus networks are necessarily more spread out, and nodes more disseminated. In this paper we adjust the method from Coll Ardanuy and Sporleder (2014) [3] to suit historical news stories and adapt it to meet the needs of historians.

3 The Data

From the large repository of the Dutch Royal Library, we have selected three national newspapers that reflect the most relevant aspects of the Dutch political landscape. Although the concept of pillarization came under fierce scrutiny for the last two decades and the religious and socio-political stratification of Dutch society has been questioned, the press landscape remains to be seen as fragmented. Different socio-political communities had their own media outlets that — at least until the early sixties — were neglected by other groups. Here we use a distinct catholic newspaper (*De Tijd*) and a socialist one (*Het Vrije Volk*). Results from these newspapers will be compared and contrasted with articles from *De Telegraaf*, the largest Dutch daily that had no formal political and religious affiliation [23] [30].

We focus in this case study in the first post-war decade, the period between 1945 and 1955, in which the idea of a European Union started to take shape. In order to limit the data to the pieces of news relevant to European integration, we opted to consider only articles in which the words *'Europa'* (Europe), *'Europese'* (European) or *'Europeaan'* (European) appear. This is a wide search, but it is a first step in order to reduce the corpus to the topic in which we are interested. Only articles with a high OCR confidence were considered. After filtering the articles, our dataset consisted of 2327 articles from *De Tijd*, 2663 articles from *Het Vrije Volk*, and 1138 articles from *De Telegraaf*.

4 The Method

In this section, we describe the method that we use to build social networks from a collection of news articles. A network consists of two main components: nodes and edges. In a social network, the nodes are the actors and the edges represent the relations between them. We explain in subsection 4.1 our method for obtaining the nodes of the network. In subsection 4.2 we describe how we choose to define the edges between the nodes. The creation of the network is detailed in subsection 4.3.

4.1 Obtaining the Nodes

Human Name Recognition. A social network is a structure that captures the relations between a set of actors. Thus, the first step to the creation of a social

network must necessarily be the extraction of human names from raw text. To that end, we use the `Stanford Named Entity Recognizer`.[1] We used training data for Dutch from the CoNLL-2002 shared task.[2] With the assumption that the more data, the better the entity recognition will be, we concatenated the training file together with the two test files in order to have more training data. The resulting training file consisted of 309683 tokens, 3032 of which were person names. Our training data is extracted from newspapers and, as we have already mentioned previously, we work on Dutch news text from the 1945–1955 decade. Considering that Dutch language has not changed significantly since, we expect the recognizer to work on our data as well as on modern-day data.

In order to enhance the performance of the named entity recognition module, we have applied some hand-made filtering steps, based on observation. We have realized that, on many occasions, newswire text introduces a person name by its description. In this way, it is normal to find occurrences in news text such as *'de 63-jarige Frank Donoghue'*[3], *'de kapitein Ben Shaw'*[4], or *'de 21-jarige pianist Theo'*.[5] Such linguistic cues are very reliable, since we can expect that most of the times a capitalized word following an age or a title/profession will be a person name. Two rules have been created to capture the age and the title/profession. The first one, very simple, captures every sequence of capitalized words (including initials and middle words such as *'van der'* or *'v.d.'*, typical of Dutch person names) following the expression XX-jarige, in which XX is a number (expressed numerically or alphabetically) and in which the dash is optional. To capture the title or profession, we have relied on a list of professions from the Wikipedia[6] as well as on a list of titles and professions automatically retrieved from our text, by capturing the uncapitalized word between an age expression and a capitalized word. We ended up with a list of 1650 titles or professions that are an indication that the next capitalized word will be a person name. These rules have been combined to also find entities introduced by both the age and the profession/title.

Newspapers tend to personalize institutions or organizations such as political entitites, which is the reason why, unlike in other domains such as literature, we could not rely on verbs of utterance to identify human names unequivocally. The only verbs of utterance that we have included in our filtering are mostly those that describe the manner in which something is said or which describe an action that cannot be (or is usually not) metaphorized, such as *think*, *laugh*, or *cry*, both in 3^{rd} person present and past tense. We had to disregard verbs such as *admit*, *answer* or *maintain* since most of the times the subject of these verbs are institutions or organizations. Evaluated on a small sample of 10 articles, the linguistic patterns improve the f-score of the entity recognizer from 0.70 to 0.76.

[1] http://nlp.stanford.edu/software/CRF-NER.shtml
[2] http://www.cnts.ua.ac.be/conll2002/ner/
[3] Translation: 'the 63-year-old Frank Donoghue'.
[4] Translation: 'the captain Ben Shaw'.
[5] Translation: 'the 21-year-old pianist Theo'.
[6] http://nl.wikipedia.org/wiki/Lijst_van_beroepen

Co-reference Resolution. For each human name that we identify, we keep information such as the age, the profession or title in these cases in which there is such an information. For example, if a text talks about *de 20-jarige schipper Pim de Boer*, we keep as attributes of 'Pim de Boer' his title or profession ('shipper' in this case, but we could also have some less informative title such as 'Mr.') and the possible two years in which he was born (calculated by substracting his age from the year of publication of the article).

We resolve co-reference resolution per document by string matching. We assume that two identical surface forms from the same article will refer to the same person, unless it is indicated by means of the age, title or profession. When there is a contradiction of age between two surface forms, each surface form is supposed to be an entity on its own. When the contradiction is on the title, each surface form is supposed to be a different entity only when the titles indicate two different genders (such as *'heer de Muis'*[7] and *'vrouw de Muis'*[8]). In any other case, all identical surface forms are considered to correspond to one only entity. From each article, we extract a list of human entities, each of which having the following three attributes, which might be empty if the information is unknown:

1. The list of alternative names in which this entity may be referred, including initials instead of first and middle names, contractions for particles such as 'van', etc.
2. The year in which the person was born.
3. The list of titles or professions which precede the name in the text.

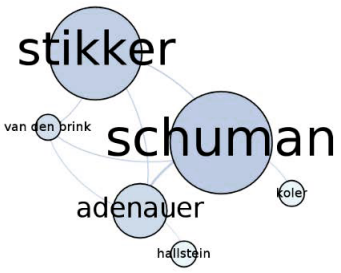
Co-reference resolution is performed in the whole dataset by string matching. We do not perform disambiguation of names. However, we give the possibility to the historian to check the list of titles and professions extracted for each entity so that we can manually correct whenever two different persons have been put together as one only entity. It is then up to the historian to decide if, for example, a farmer called Robert Schuman is the same person as the politician called Robert Schuman.

4.2 Establishing the Relations

Once we have found the nodes of the network, we need to define what kind of relation we want to draw between them. In our case we created an undirected graph based on the co-occurrence of nodes in each article. In other words, each pair of nodes is linked in our network if they co-occur in the same news article. Our network is weighted, so the more two entities interact throughout the collection, the stronger the relation between them will be. In an edge attribute we keep the list of articles in which both nodes of the edge co-occur. With this measure, we allow the historian to go back to the source files in which each two nodes co-occur. Each edge in the network has two more attributes apart from the

[7] Translation: 'Mr. de Muis'.
[8] Translation: 'Mrs. de Muis'.

weight and list of files: the list of the most common words (stopwords removed) of the articles in which both entities are mentioned and the list of the most relevant words using the tf-idf weighting for the concatenation of documents in which both nodes appear. An example of extracted attributes for a pair of nodes can be seen in Fig. 1.



Attributes of the node *Adenauer*
aka: Konrad Adenauer, Adenauers
born: unknown
title: heer, Bondskanselier, heren, kanselier, dr.

Attributes of the node *Schuman*
aka: Robert Schuman, Schumans, R. Schuman
born: unknown
title: heer, minister

Attributes of the edge *Adenauer-Schuman*
tfidf: sociaal-democraten, regeringsverklaring, Woensdagavond, Bondsdag, nota, Saarland, Commissie, goedgekeurd, Adenauer, plan-Schuman, Hoge, minister, Buitenlandse, verdrag, Zaken, gezet, gelijkheid, Kanselier
counts: minister, buitenlandse, europese, plan-schuman, zaken sociaal-democraten, acht, adenauer, bondsdag, commissie, duitse, europa, gemaakt, goedgekeurd, grote, hoge, kanselier, lang
co-occurring files: id0243.txt, id0566.txt, id0568.txt, etc.

Fig. 1. On the left, a simplified representation of a fragment from the network extracted from *De Tijd*, year 1951. On the right, attributes of the nodes *Adenauer* and *Schuman* and of the edge *Adenauer-Schuman*.

4.3 Building the Network

Since we are interested in the development of the network throughout history, we create dynamic networks, i. e., the succession of yearly static networks, for each collection. The python library `Networkx`[9] is used to construct the networks and the network analysis software `Gephi`[10] is used to visualize them. Both are open-source tools with an extensive functionality.

5 Analysis

The resulting social networks offer some expected results that prove to a certain degree the reliability of our approach, such as the important presence in the network of personae such as Robert Schuman, Dirk Stikker, Ernest Bevin, Konrad Adenauer, Winston Churchill, Georges Bidault, Willem Drees, Alcide de Gasperi, or Jean Monnet. It could also be expected that a socialist newspaper such as *Het Vrije Volk* would give more weight to local stories and relatively less weight to big names than other newspapers. Indeed, in our graphs, the weight

[9] http://networkx.github.io/
[10] http://gephi.org/

of the 10 most common nodes in *De Tijd* is 16% of the total weight, in *De Telegraaf* is 12%, and in *Het Vrije Volk* is 10%. While expected results are useful to understand that the method works and is able to predict correctly certain facts, it is when unexpected results come out that our approach is most interesting. Unexpected results are potential hypotheses that defy official history. It is then the task of the historian to verify, by looking at the pieces of news selected by our method, whether there is some truth in the information yielded by the network.

Looking at the results of our digital analysis, the importance of politics immediately stands out. Central actors in all networks continuously were politicians. This seems to be less surprising as it is. Milward [14] and Moravcsik [15] in their work on European integration named economic self-interest of the state the most important driver in the integration process. Combining our extracted networks with the word counts and tf-idf analysis, we come to the conclusion that early integration is better framed as essentially a political process. More than a cooperation of states, it was the creation of a new political sphere with its own vocabulary, central actors and use of concepts [27] [8].

The centrality of American politicians merits particular attention. In the literature, the importance of America as an actor in the early integration process is emphasised only for the late forties. After the presentation of the Schuman plan, the basis for the foundation of the Coal and Steel Union, in May 1950 the integration process more and more became truly a European matter [10]. But our networks clearly show that, at least in public discourse, America remained to be seen as an important actor. More specific research could be done here. Another issue raised is the concurrent ongoing emphasis on early integration being a technocratic process with a low political profile. As Mark Mazower concluded, the Second World War had left people with a deep antipathy towards ideological politics, which was reflected by mainstream politics steering away from polarized attitudes in favour of compromise. The European project after the presentation of the Schuman plan was one of the most important examples to support such a claim [11]. Recent studies however have reemphasized the role of ideology in early integration history and our research seems to prove them right, although the width and nature of the used material urges some caution here. Nonetheless, it is worth noticing that words like *'gemeenschapszin'* (sense of community) and *'solidariteit'* (solidarity) seemingly played a role in public discourse.

This corresponds with recent studies being done on political parties and civil society organisations that became laboratories of community in postwar Europe. After the Second World War, all over Europe new ideas on community arose out of the desire for stability, prosperity and welfare after years of devastating violence. Initiatives to reconcile and reunite European citizens reflected these debates. In that sense, European integration was merely a peace process [24]. Our material can be used to support such a claim, for instance by looking at the minor but significant differences in actors and vocabulary between the three newspapers. Seemingly, different moral communities formed different epistemic communities that supported or criticized the work of political leaders.

6 Conclusion

The central objective of this paper was to see how computational techniques could strengthen the empirical foundations of new European integration history. In the growing field of digital humanities, many voices have expressed a fear of a decline in the role of interpretative close reading [19] [25]. Although we do not completely share this fear, we do see the importance of combining different research strategies. In our paper, we have shown that using network extraction raises new questions on early European integration and suggests an outline for new and more refined research. The use of large digitised repositories and digital strategies to extract, select, analyse and read the material can be a potential way to overcome a problem immanent to European integration history: it is transnational, multilingual and ramified. Especially now that public discourse seems to have become the focal point in new historiography, digital search tools and data mining techniques can greatly further the scope of inquiry, as long as we remain critical towards some frames of big humanities. We have presented our approach as a showcase for digital humanities in the field of European integration history with the hope that it can become a stepping-stone for further research.

Acknowledgments. This project was funded as part of the HERA programme.

References

1. Bamman, D., O'Connor, B., Smith, N.A.: Learning latent personas of film characters. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp. 352–361 (2013)
2. de Bruin, R.: Elastisch Europa. De integratie van Europa en de Nederlandse politiek, 1947–1968. Wereldbibliotheek, Amsterdam (2014)
3. Coll Ardanuy, M., Sporleder, C.: Structure-based clustering of novels. In: Third Workshop on Computational Linguistics for Literature at EACL 2014, pp. 31–39. Gothenburg, Sweden (2014)
4. Elson, D.K., Dames, N., McKeown, K.R.: Extracting social networks from literary fiction. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 138–147 (2010)
5. de Roode, S.L.R.: Seeing Europe Through the Nation. TheRole of National Selfimages in the Perception of European Integration inthe English, German, and Dutch Press in the 1950s and 1990s. Steiner, Stuttgart (2012)
6. Hooghe, L., Marks, G.: Europe's blues. Theoreticalsoul-searching after the rejection of the european constitution. Political Science & Politics **39**(2), 247–250 (2006)
7. Jackson, C.A.: Using Social Network Analysis to Reveal Unseen Relationships in Medieval Scotland. In: Digital Humanities Conference, Lausanne (2014)
8. Kaiser, W.: Transnational western europe since 1945. Integration as political society formation. In: Kaiser, W., Starie, P. (eds.) Transnational European Union. Towards a Common Political Space, pp. 17–35. Routledge, London (2005)
9. Kaiser, W.: Transnational networks in european governance. The informal politics of integration. In: Kaiser, W., Leucht, B., Rasmussen, M. (eds.) The History of the European Union. Origins of a Trans- and Supranational Polity 1950–1972, pp. 12–33. Routledge, New York/London (2009)

10. Lundestad, G.: "Empire" by Integration: The United States and European Integration, 1945–1997. Oxford University Press, Oxford (1998)
11. Mazower, M.: Dark Continent. Europe's Twentieth Century. Vintage Books, London (1998)
12. Medrano, J.D.: Framing Europe. Spain, and the United Kingdom. Princeton University Press, Princeton, Attitudes to European Integration in Germany (2003)
13. Meyer, J.H.: Tracing the European Public Sphere. A Comparative Analysis of British, French and German Quality Newspaper Coverage of European Summits (1969–1991). Steiner, Stuttgart (2010)
14. Milward, A.S.: The European Rescue of the Nation-state. Routledge, London (1992)
15. Moravcsik, A.: The Choice for Europe. Social Purpose and State Power from Messina to Maastricht. Cornell University Press, London/New York (1998)
16. Moravcsik, A.: What Can we Learn from the Collapse of the European Constitutional Project? Politische Vierteljahresschrift **47**(2), 219–241 (2006)
17. Oelke, D., Kokkinakis, D., Malm, M.: Advanced visual analytics methods for literature analysis. In: Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) at EACL 2012 Workshop. Avignon, France (2012)
18. Padgett, J.F., Ansell, C.K.: Robust action and the rise of the Medici, 1400–1434. American Journal of Sociology **98**(6), 1259–1319 (1993)
19. Piersma, H., Ribbens, K.: Digital Historical Research. Context, Concepts and the Need for Reflection. BMGN - Low Countries Historical. Review **128**(4), 78–102 (2013)
20. Rochat, Y., Fournier, M., Mazzei, A., Kaplan, F.: A network analysis approach of the venetian incanto system. In: Digital Humanities Conference, Lausanne (2014)
21. Seidel, K.: From pioneer work to refinement. publication trends. In: European Union History: Themes and Debates, pp. 26–44. Palgrave, Basingstoke (2010)
22. Trenz, H.J.: Media: The unknown player in european integration. In: Media. Democracy and European Culture, pp. 49–64. Intellect, Bristol (2008)
23. van Dam, P.: Staat van verzuiling. Over een Nederlandse mythe. Wereldbibliotheek, Amsterdam (2011)
24. van den Bos, M.: Mensen van goede wil. Pax Christi Nederland (forthcoming)
25. van Eijnatten, J., Pieters, T., Verheul, J.: Big Data for Global History: The Transformative Promise of Digital Humanities. BMGN - Low Countries Historical Review **128**(4), 55–77 (2013)
26. van Middelaar, L.: Telling Another Story of Europe. A Reply in Favour of Politics. BMGN - Low Countries Historical. Review **125**(4), 82–89 (2010)
27. van Middelaar, L.: The Passage to Europe: How a Continent Became a Union. Yale University Press, London (2013)
28. Wendt, A.: Social Theory of International Politics. Cambridge University Press, Cambridge (1999)
29. Wieneke, L., Düring, M., Sillaume, G., Lallemand, C., Croce, V., Lazzaro, M., Nucci, F.S., Pasini, C., Fraternali, P., Tagliasacchi, M., Melenhorst, M., Novak, J., Micheel, I., Harloff, E., Garcia Moron, J.: Building the social graph of the history of european integration. In: HistoInformatics Workshop at SocInfo 2013, Kyoto (2013)
30. Wolf, M.: Het geheim van De Telegraaf: geschiedenis van een krant. Boom, Amsterdam (2009)