Historical Event Search in Digital Heritage

Studying Commemorative Practices in Diachronic Corpora

Kaspar Beelen
University of Amsterdam
k.beelen@uva.nl

Alex Olieman
University of Amsterdam
olieman@uva.nl

Jaap Kamps
University of Amsterdam
kamps@uva.nl

ABSTRACT

The past is an unattainable country. All access to it is mediated by myriad artifacts. To make sense of this morass of information, historians invented concepts (such as "the French Revolution" or "the Dutch Golden Age") that bind an otherwise unconnected set of entities together. These historical periods, however, remain constructions and their meaning a moving target. In this paper, we outline a search interface that allows researchers to explore mentions to past events or periods in digital heritage. We show how semantically-enhanced search enables users to retrieve information related to complex concepts. After introducing the general architecture and the interface, we showcase it by elaborating on one pilot study, targeting the Golden Age in contemporary Dutch parliamentary discourse.

Keywords: Colligatory Concepts, Semantically-Enhanced Search, Interactive Information Retrieval, Corpus Selection, Digital Humanities

1 INTRODUCTION

The past is not a foreign, but an unattainable country. All access to it is mediated via myriad artifacts, which historians attempt to weave together in convincing narratives [15]. To this end, scholars tailor concepts that bind an otherwise heterogeneous set of entities and events into coherent historical stories. Many of these concepts subsequently outgrow their professional origins and nestle themselves in popular discourse—the "Renaissance" [2] serves as an example of such an invention. Scrutinizing the content and structure of these concepts—the way historical periods are represented—uncovers how societies deal with their past. Even though the study of memory forms a crucial and popular topic in the humanities [5, 7], scholars often find it problematic to find and identify historical references in large corpora. In this paper, we, therefore, demonstrate how semantically-enhanced search helps historians tracing references to complex historical constructs in digital heritage.

From an Information Retrieval perspective, this amounts to a daunting task: finding documents that refer to fragmented and heterogeneous concepts is fraught with multiple methodological and philosophical hurdles. How, for example, can a machine retrieve all documents related to the "French Revolution"? To contextualize the problem, Shaw [11] introduced the term "colligatory concept" to Information Science. The notion of colligation originated from Whewell [14], and was applied to the philosophy of history by Walsh [13]. In the latter discipline, colligatory concepts are inventions made by historians that group together various facts, events and persons, inferred through an inquiry of the past. The concepts historians forge are an attempt to make the past understandable by imposing mental constructs on the data. Shaw [11] distinguished historical periods as a prevalent form of colligation, since they group various entities—ranging from persons (e.g. Robespierre, in the case of the French Revolution) over locations (Bastille) to time (1789)—under one header. These representations are not just complex, but also unstable since their content varies depending on the perspective of the narrator. This paper introduces a framework for searching and modeling colligatory concepts in digital heritage. It attempts to tackle the technical as well as (part of) the philosophical hurdles.

The remainder of the paper is structured as follows: Firstly, we describe WideNet, a novel (re)search interface that is designed to explore historical periods in diachronic corpora. Secondly, we discuss a specific pilot study—the Golden Age in contemporary Dutch Parliamentary discourse—to demonstrate how semantically-enhanced search supports historians in studying how the past is remembered.

2 SEARCHING FOR COLLIGATORY CONCEPTS IN PARLIAMENTARY SPEECH

WideNet builds on a semantically enriched version of the Dutch parliamentary proceedings: the "verbatim" record of all debates in the *Staten Generaal*). These discussions touch on almost every issue that moved Dutch public opinion during the last two centuries. Despite its centrality in the political landscape, the proceedings' unwieldy size made it difficult for historians to explore. The existing index is rather limited in scope, which makes searching for infrequent and complex items a laborious, if not impossible, task. Digitization and enrichment have unlocked this resource in novel ways. WideNet, we believe, is a valuable addition to this trend, since it accommodates a growing need for complex search systems in the Digital Humanities. The application of semantically-enhanced search to large digitized text collections was first proposed by Hinze et al. [3]. Semantically-enhanced search aims to overcome the gap between the research questions and methods of the humanities and full-text (lexicographic) search.

2.1 Offline Processing

To prepare the corpus for semantically-enhanced search, all the documents were processed by a semantic annotation system, which linked concepts and entities in the text to a Knowledge Base (KB) [3]. In previous work [8] we generated entity links for a collection of

SEMANTiCS 2017 workshop proceedings: EVENTS
September 11-14, 2017, Amsterdam, Netherlands

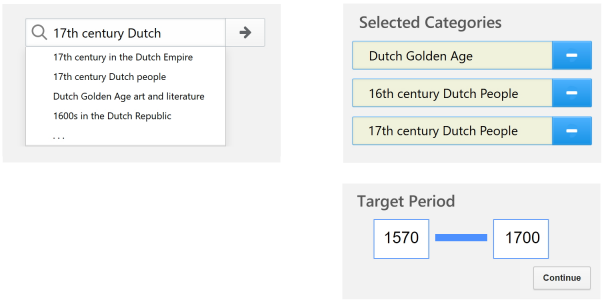

Figure 1: Initial query specification in WideNet.

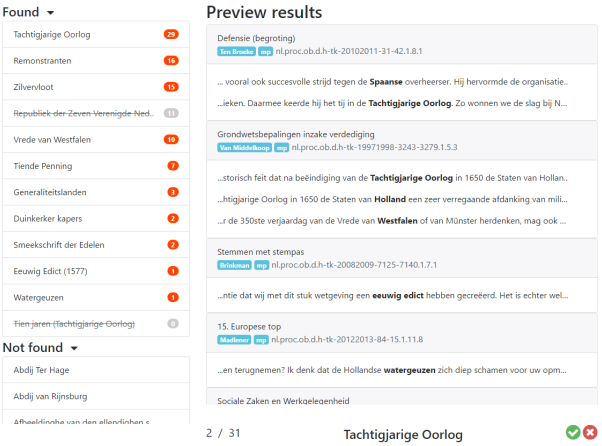

Figure 2: Assessing the relevance of categories and entities.

parliamentary proceedings using DBpedia Spotlight. The DBpedia Spotlight annotations, which obtain an estimated precision of 0.69 and recall of 0.40 [8], were added to existing search indexes as additional (nested) fields.

To map abstract concepts like "Dutch Golden Age" to more specific concepts and entities, we extracted a subgraph of DBpedia into a property graph database (see [10]). The category network is used at runtime to select potentially relevant entities given a root category, by traversing `dct:subject`[1] and `skos:broader` relations in reverse direction. Our proof-of-concept makes use of DBpedia, but any KB that conforms to the SKOS ontology can in principle be loaded. Finally, the system needs access to coarse temporal clues about entities. Because DBpedia does not provide this data reliably across entity types, we extract mentioned years from the `rdfs:comment` values of DBpedia resources with a simple regular expression, and add them to the graph.

2.2 Search Interface Design

The user-interface guides scholars through three phases: (1) selection of the root category, (2) assessment of the categories' and entities' relevance (3) close-reading of selected documents. In the first step, the user selects a root category from a typeahead search box (see Figure 1), and demarcates the query by selecting a time

[1]For namespace prefixes, see https://dbpedia.org/sparql?nsdecl.

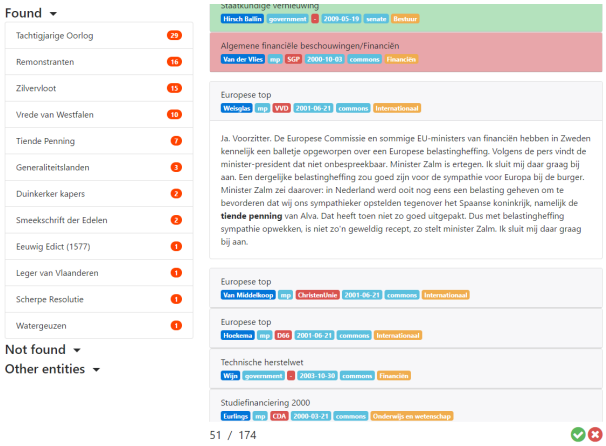

Figure 3: A closer look at the retrieved documents.

period, which is used to prune the underlying entities of the selected categories. WideNet then retrieves the network of narrower categories for each selected root category and collects the contained entities as potentially relevant query components. Behind the scenes, each entity is compared with the target period (i.e. time interval), and is considered to be outright relevant to the period, or not, or a borderline case, or as lacking temporal clues altogether. In the current implementation this classification is achieved with simple rules, based on the features: 'fraction of years within period,' 'fraction of intervals that overlap with the period,' and 'has at least one year in period.' The system uses chronological information to deselect (sub)categories where more than half of the dated member entities are out-of-period. Subsequently, the user assesses which of the retrieved subcategories actually contain entities that lead to relevant results (see Figure 2). This step was motivated by the following observations: Firstly: many of entities returned by Spotlight were simply incorrect and therefore should be discarded easily. Secondly: as we lack a clear ontology of our target—as stated earlier, colligatory concepts are by definition dependent on the perspective of the user—the interface shows a wide range of *potentially* relevant entities, but defers the actual selection to the scholar, who is ultimately responsible for judging the "aboutness" of a reference [11]. The interface facilitates this task by showing, per subcategory, which entities are mentioned in the corpus, and how frequently. Thirdly: to make users aware of the "silences" with respect to the queried concepts, we also listed all entities that were searched for but did *not* occur. After selecting relevant categories of entities, the WideNet interface allows further inspection in the form of close-reading, as shown in Figure 3. This enables the users to compile a corpus of relevant documents which may be saved and exported. Moreover, the user can examine the selected documents in relation to their metadata, e.g. look for saliency by plotting the annotations over time, or study bias by aggregating the results by political party (see Figure 4).

3 PILOT STUDY: MINING THE GOLDEN AGE

3.1 Background and Motivation

The study of "memory"—the diverse ways through which historical events reverberate over time—has attracted the interest of historians and other scholars working in the humanities[5]. Especially in the study of nations and nationalism, the past weighs heavily [7]. A distinct discipline even emerged, called "Imagology", which focuses on the critical analysis of national stereotypes and their historical origins [6]. We applied WideNet to analyze the changing face of Dutch nationalism from the late 1990s to the present via a dissection of narratives related to the Golden Age. This study is situated within a recent stream of literature that looked at the changing discourse on "Dutchness" (for a critical review see [9]). In broad lines, it argues that the Netherlands experienced a rapid transition from a "thin" to a "thick" conception of national identity. The rather abrupt return to nationalism saw a procedural and heterogeneous perception of Dutchness being substituted by a culturalist, homogenizing version [12]. To assess the impact of the past on the construction of Dutch national identity [16], we scrutinized the speeches of parliamentarians for references to one of the most celebrated eras in Dutch history: the so-called "Golden Age." Since the 19th century, this period—a fine example of a colligatory concept—has served as a benchmark ("ijkpunt") of national identity [4]. The Golden Age has played a crucial role in defining the Dutch national "We". Events and individuals from this era came to symbolize national characteristics: the bravery of the "Geuzen," the enlightened thinking of Spinoza, all these historical themes composed a rich resource for identity construction.

3.2 Data Preparation and Selection

Traditionally, scholars in the humanities tend to base their arguments on a small set of carefully selected and curated material. Instead of building conclusions on small datasets, the digital turn has enabled historians to "holistically assess the typicality, scope, and power of key issues" [1]. In this respect, WideNet assists researchers with the digital exploration of complex concepts in large diachronic corpora. It introduces a more data-driven approach to corpus selection in the humanities. The case study draws on a digitized version of the Dutch parliamentary proceedings—more specifically the debates of the Lower and Upper Houses between 1995 and 2014—which were richly adorned with semantic annotations as part the PoliticalMashup project (which later was continued under the Dilipad acronym).[2]

3.3 Analysis and Reflection

We started by sifting through all the entities WideNet returned as related to the Golden Age, using the interface shown in Figure 2.[3] Having a thematic overview of all entities facilitated this task for three reasons. Firstly: as entity-linking remains an error-prone process, many of the items found by WideNet turned out to be irrelevant. For example the historical person of Michiel de Ruyter was often confused with a hospital named after him. Judging whether an entity relates to the topic at hand is, as we argued previously, ultimately vested in the historian who models the period under investigation. In line with the description of colligatory concepts, the design of the WideNet interface facilitates this task by actively engaging the user to refine the search results and define the target query.

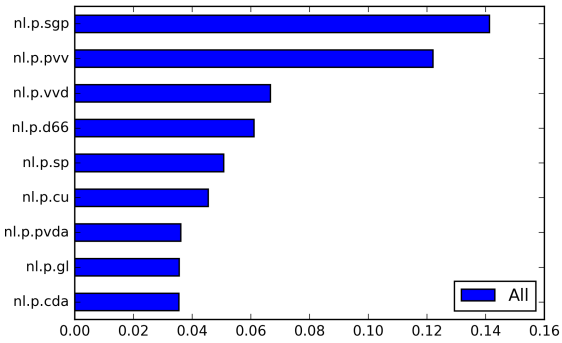

Figure 4: References by Party.

Secondly: as the overview groups all entities by their subcategory, we can easily identify the distinct topics and spheres of identity construction. We labeled the found entities as either belonging to the economical (i.e. Dutch East India Company), cultural (i.e. Spinoza), or the political sphere (i.e. Eighty Years' War). The coarse-grained subcategory-wise exploration of the retrieved entities foregrounded the diversity of the search result. It, thereby, also made long-tail entities clearly visible: grouping entities by subcategory prohibits the rare ones from drowning in a morass of highly frequent (but potentially irrelevant) results. For example: mentions of the architectural artifacts of the seventeenth-century appeared very often, but did not co-occur with any relevant event or person (and were thus conveniently discarded with one click), while references to the Eighty Years' War were scarce but highly informative.

Thirdly: besides thematically grouping the references, WideNet allows the user to aggregate the selected entities by their metadata fields. For our purposes, we were particularly interested in the distribution of these mentions over time and party. From a temporal perspective, the number of Golden Age references showed a slight increase after 2002, but not drastically (result not reproduced here). Aggregating the found references by party revealed deeper-rooted discrepancies in the engagement with the past. The discourse of the conservative and populist right showed a greater tendency to invigorate their nationalistic appeals with historical references. As shown in Figure 4 the Reformed Party and the Party for Freedom (PVV), were proportionally the most active in this respect, followed by the right-liberal People's Party for Freedom and Democracy (VVD). The fact that most of these mentions stemmed from the right, suggests that the Golden Age functioned as an instrument for forging a more exclusive, culturalist understanding of Dutchness. This applied even more to political references, which almost exclusively circulated among those seated at the right. Mentions of

[2] see http://politicalmashup.nl/ and https://dilipad.history.ac.uk/
[3] The original data are available on https://widenet.e.hum.uva.nl/preview/ge/

economic entities showed a more balanced distribution, skewing even a bit towards the left (results not reproduced here).

By aggregating the selected speeches, WideNet enables researchers to conveniently map their data along different axes. But what do these aggregated differences actually mean? How, exactly, do these speeches enact of identities? What are the specific linguistic instruments that "bound and bond people" into distinctive groups? To answer these questions, we studied the concrete, fine-grained, mechanics at play in the speeches that survived the filtering process (as shown in Figure 3). Linguistically, the use of personal pronouns, such as "we" became apparent, especially in the in the context of political entities—which mainly comprised wars and taxation, such as the Eighty Years' War and the "tiende penning," a VAT on movables, introduced by the Duke of Alva. For example, Verheijen a MP of the right-liberal VVD, evoked these events when commenting on proposals which would give Europe more say in national taxation matters. To support his argument he asserted that "we waged 80 years of war against Spain to obtain our independence." In this passage, Verheijen inserts a transhistorical "we" in his speech, he amalgamates the Dutch who fought Spain with those of today, who still fiercely resist infringement on their national independence.

Generally, a fine-grained linguistic analysis, demonstrated that MPs on the right exhibited greater intimacy with the past: more often did they invoke the "we" as a homogeneous national actor, or used cognitive verbs to suggest direct access to the minds of illustrious Dutchmen from the past. Another example is Madlener of Wilders' PVV, who insisted that "we won against Spain"—again a reference to the Eighty Years' War during the seventeenth century. Moreover, he exclaims: "I think the Sea Beggars are deeply ashamed of your remark", thereby rejecting the claims of an opponent by straightforwardly probing and exposing the minds of actors from the distant past.

While the memory of political events was dominated by the right, the discussion about the economic ramifications of the Golden Age was more evenly distributed among left and right—but nonetheless contested. The interpretation of the Dutch colonial heritage and its trade practices figured here as the main bone of contention. This debate was largely sparked by a remark of then-prime-minister Jan Peter Balkenende, who urged the Dutch to embrace their "VOC-mentality," which he characterized as a tradition of risk-taking and brave, global entrepreneurship. The depiction of the VOC as the flagship of Dutch capitalism—based on the values "Freedom, Entrepreneurship, and Competition", according to Ten Broeke of the VVD—bounced against a wall of skepticism raised by left-wing MPs, who mostly emphasized the exploitative practices of the colonial past. Vendrik of GreenLeft, for example, posits that the VOC mentality actually means "becoming rich at the expense of others", while Irrgang, a member of the Socialist Party, equates the VOC—and the underlying mentality—with "nothing more than colonial plundering, the creation of monopolies."

4 CONCLUSION

The Golden Age serves just as one example of how WideNet supports the scholar to study historical periods in large corpora. We have shown how the WideNet interface enables historians to explore colligatory concepts in different stages: First, the user selects the query from an existing Knowledge Base, after which WideNet gathers all documents that contain references to the different aspects of the historical period under investigation. The user is then presented with an overview of the found subcategories and their associated entities. At this stage the user decides which of the found events and persons are relevant and should be retained and subjected to close reading. Eventually, the findings can be aggregated and plotted along various dimension such as time or party.

During the search process, WideNet logs all activity and thereby collects a fair amount of data. Besides the initially selected target categories and time period, it stores each decision about the relevance of the entities, as well the individual documents. In other words, WideNet gathers how *users* model historical events. These data are expected to be valuable for two purposes. Firstly, as a gold standard for the (temporal) pruning of raw category trees. Collecting user data enables the system to automatically deselect irrelevant entities (which would still be reversible by the user). Secondly, logging the decisions made by many users might foreground academic or societal disputes about certain events in the past. Besides providing a useful front-end for historical research, the interface's back-end could return valuable data for humanities scholars, as it may expose different understandings of history that circulate in society.

REFERENCES

[1] Luke Blaxill. 2013. Quantifying the language of British politics, 1880-1910. *Historical Research* 86, 232 (2013), 313–341. https://doi.org/10.1111/1468-2281.12011

[2] Jacob Burckhardt. 1914. *The civilization of the Renaissance in Italy.* G. Allen & Unwin, Limited.

[3] Annika Hinze, Craig Taube-Schock, David Bainbridge, Rangi Matamua, and J. Stephen Downie. 2015. Improving Access to Large-scale Digital Libraries Through Semantic-enhanced Search and Disambiguation. In *Proceedings of the 15th ACM/IEEE-CE on Joint Conference on Digital Libraries - JCDL '15.* 147–156. https://doi.org/10.1145/2756406.2756920

[4] Lotte Jensen. 2012. De Gouden Eeuw als ijkpunt van de nationale identiteit. Het beeld van de Gouden Eeuw in verzetsliteratuur tussen 1806 en 1813. *De Zeventiende Eeuw. Cultuur in de Nederlanden in interdisciplinair perspectief* 28, 2 (2012).

[5] Wolf Kansteiner. 2002. Finding meaning in memory: A methodological critique of collective memory studies. *History and theory* 41, 2 (2002), 179–197.

[6] Joep Leerssen et al. 2007. Imagology: History and method. *Imagology: The cultural construction and literary representation of national characters* (2007), 17–32.

[7] James H Liu and Denis J Hilton. 2005. How the past weighs on the present: Social representations of history and their role in identity politics. *British Journal of Social Psychology* 44, 4 (2005), 537–556.

[8] Alex Olieman, Jaap Kamps, Maarten Marx, and Arjan Nusselder. 2015. A Hybrid Approach to Domain-Specific Entity Linking. In *Joint Proceedings of the Posters and Demos Track of 11th International Conference on Semantic Systems - SEMANTiCS2015 and 1st Workshop on Data Science: Methods, Technology and Applications (DSci15).* Vienna, 55–58. http://arxiv.org/abs/1509.01865

[9] Rogier Reekum. 2012. As nation, people and public collide: Enacting Dutchness in public discourse. *Nations and Nationalism* 18, 4 (2012), 583–602.

[10] Marko. Rodriguez. 2015. The Gremlin Graph Traversal Machine and Language. In *Proc. 15th Symposium on Database Programming Languages.* 1–10. https://doi.org/10.1145/2815072.2815073 arXiv:1508.03843

[11] Ryan Shaw. 2013. Information Organization and the Philosophy of History. *Journal of the American Society for Information Science and Technology* 64, 6 (jun 2013), 1092–1103. https://doi.org/10.1002/asi.22843

[12] Evelien Tonkens, Menno Hurenkamp, and Jan Willem Duyvendak. 2010. Culturalization of citizenship in the Netherlands. *Managing ethnic diversity after 9/11: integration, security, and civil liberties in transatlantic perspective* (2010).

[13] William H Walsh. 1942. The intelligibility of history. *Philosophy* 17, 66 (1942), 128–143.

[14] William Whewell. 1858. *Novum organon renovatum.* JW Parker and son.

[15] Hayden White. 2009. *The content of the form: Narrative discourse and historical representation.* JHU Press.

[16] Ruth Wodak. 2009. *Discursive construction of national identity.* Edinburgh University Press.