Storyteller: Visualizing Perspectives in Digital Humanities Projects

Janneke M. van der Zwaan[1(✉)], Maarten van Meersbergen[1], Antske Fokkens[2],
Serge ter Braake[3], Inger Leemans[2], Erika Kuijpers[2], Piek Vossen[2],
and Isa Maks[2]

[1] Netherlands eScience Center, Amsterdam, Netherlands
`j.vanderzwaan@esciencecenter.nl`
[2] VU University, Amsterdam, Netherlands
[3] University of Amsterdam, Amsterdam, Netherlands

Abstract. Humanities scholars agree that the visualization of their data should bring order and insight, reveal patterns and provide leads for new research questions. However, simple two-dimensional visualizations are often too static and too generic to meet these needs. Visualization tools for the humanities should be able to deal with the observer dependency, heterogeneity, uncertainty and provenance of data and the complexity of humanities research questions. They should furthermore offer scholars the opportunity to interactively manipulate their data sets and queries. In this paper, we introduce Storyteller, an open source visualization tool designed to interactively explore complex data sets for the humanities. We present the tool, and demonstrate its applicability in three very different humanities projects.

Keywords: Visualizations · Heterogeneous data · NLP · History · Storylines

1 Introduction

Visualizations in digital humanities often represent the results of complex algorithms running on 'big data' from the humanities domain. Humanities scholars agree that the visualization of their data is an essential instrument to communicate their research results and to evoke scholarly debates as well as public interest [5,7,10]. The disciplinary characteristics of both scholarly practices and data in the humanities, however, offer a challenge for visualization techniques. In terms of practice, the humanities have stringent requirements concerning the representation of data and their provenance as their data are usually not generated under controlled circumstances. Humanities scholars like to have direct access to their data, queries and results in order to manipulate them and discuss their interpretation and also provide this to other users. Because of the heterogeneity of humanities data, and the exploratory character of humanities research

questions, a limited corpus of texts can already generate an extensive amount of complex data [2].

In this paper, we address these challenges and present Storyteller, a visualization tool that helps to analyze complex, multilayered data. The tool allows users to interactively explore the data by adjusting the queries. In addition, Storyteller takes provenance into account by allowing users to view the data in the original context of their source. We have used Storyteller to visualize data from three different digital humanities projects. The first project is 'NewsReader'[1] which automatically detects events in news in terms of what happened, who is involved, and where and when it happened. In addition, events are aggregated into storylines to provide explanatory structures. The second project is 'BiographyNet'.[2] This project involves the representation of biographies of prominent figures in Dutch history. The third project is Embodied Emotions.[3] The goal of this project is to track changes over time in the relationship between emotions and the parts of the body that are involved in emotional experience and expression. While the data of these three projects are very different, we demonstrate that Storyteller provides a new visualization technique that offers enhanced insights for each of these projects.

This paper is organized as follows. Section 2 provides a general introduction to the specific challenges humanities research sets for visualization of digital research. In Sect. 3, we present the Storyteller tool. The three different use cases are discussed in Sect. 4. Section 5 presents our concluding remarks.

2 Humanities Data and Visualizations

Over the last decade, various scholars have stressed the necessity for digital humanities to develop its own visualizations that specifically cater for the needs of humanities projects. In a seminal article from 2011, Johanna Drucker stated that adaptations of visualization techniques from the natural and social sciences do not meet the needs of humanities scholars, because they simplify research data and inherently assume knowledge to be observer-independent, certain and factual. Drucker stresses that, as opposed to most scientific research where data is regarded as given, most humanities data should in fact be read as 'capta': selected and constructed [3]. This implies that data visualizations should always provide contextualization through several layers of information. Humanities scholars therefore face problems with current data visualization techniques on various levels:

1. **Multiple entities over multiple dimensions**
   Phenomena in the world of humanities research "do not lend themselves to representation within bounded, carefully delimited parameters" [4, p.247].

---

[1] http://www.newsreader-project.eu/.
[2] http://www.biographynet.nl.
[3] https://www.esciencecenter.nl/project/from-sentiment-mining-to-mining-embodied-emotions.

Single value systems are often too reductive to fit humanities requirements. Humanities research typically aims to interrelate multiple entities. Scholars will try, for instance, to map multiple events, investigate locations, and participants to these events, and plot them over time. Humanities visualizations should be able to interrelate these entities and plot them over third (space) and fourth (time) dimensions.

2. **Valuation - interrelation**
Entities and events may vary in value or weight and these in turn may differ depending on the research question. Visualizations should be able to reflect these differences. This can be done on various levels, for instance, by assessing the sentiment events generate, the number of mentions they get or by observing co-reference: is a single event mentioned in multiple sources?

3. **Provenance**
Humanities scholars have high standards for the accountability of the provenance of data: much of this data consist of original (textual) material that needs critical consideration, weighing and positioning, in order to validate the relevance of the presented results [8]. Visualizations should therefore represent data provenance. Often, this entails presenting text in visualizations, instead of just quantitative data.

4. **User dependency and perspectives**
Humanities scholars stress the fact that visualizations are not endpoints in the research cycle (a static, graphic presentation of the results), but one of the many steps in the hermeneutic cycle that can help to interpret the data, leading up to a next phase of research [5]. Furthermore, in humanities research "facts" are often contested. Visualizations should therefore be able to map multiple perspectives on the same topic or the same event. Instead of pretending to present final results, humanities visualizations should account for the uncertain and interpretative elements inherent in the data representation. Visualizations must offer researchers the possibility of manipulating the selection of data or the queries that address them in order to try and analyze different perspectives [3, p.20:51].

Visualizations for humanities research should be able to display the inherent ambiguity and complexity of 'capta', while building the co-dependent relation between observer and research data into an interactive, interpretative visualization model. Storyteller takes all these facets into account in an attempt to provide a visualization that brings new insights in humanities data.

3 Storyteller

Storyteller visualizes events in humanities data sets as structured stories, as defined in [14]. Following van den Akker et al., the tool shows interlinked actor-centric and event-centric stories that are extracted from text [1]. The properties of the data determine what are actors (participants) and events. In this section, we explain the visualizations in general terms. Section 4 discusses the data set-specific uses.

Figure 1 shows a mock-up of Storyteller.[4] The tool visualizes connections between *events* that happen over time and *participants* in these events. Events are divided into *groups*, based on data set-specific properties. Each event is associated with a *date* and a *climax score* that indicates the importance of the event. The visualization consists of three linked views.

The first view consists of a horizontal bar chart and a co-participation graph. The bar chart lists all participants in the data. The length of the bars correspond to the number of times a participant is mentioned in the source text(s). The co-participation graph shows events as nodes on a timeline. Participants are represented by lines. If multiple participants participate in the same event, their lines intersect at the node representing this event. Hovering the mouse cursor over an event shows further details.
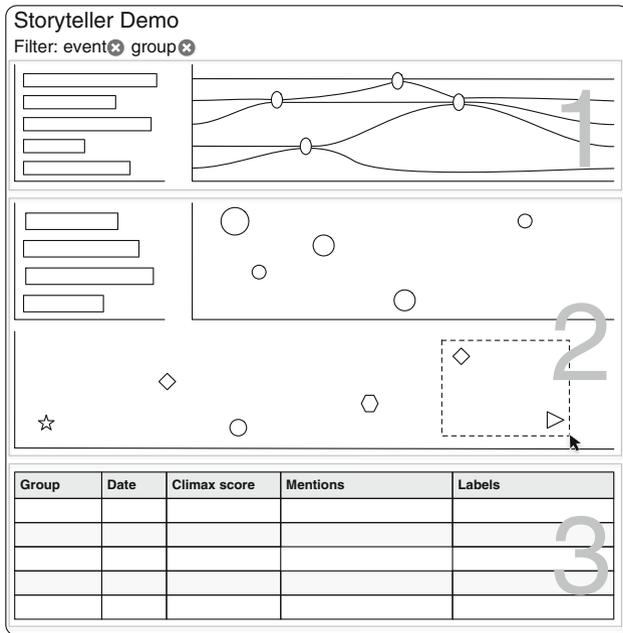

Fig. 1. Mock-up of Storyteller.

The second view consists of a horizontal bar chart and two time-based plots. The horizontal bar chart lists the groups that occur in the data set. The length of the bars indicate the number of events in the group. For each group, the top most plot displays a sequence of time-ordered events (approximating a storyline). The

---

[4] For a better understanding of the Storyteller visualizations, the online demos can be found at http://nlesc.github.io/UncertaintyVisualization/ (NewsReader), http://nlesc.github.io/BiographyNet/ (BiographyNet), and http://nlesc.github.io/EmbodiedEmotions/ (Embodied Emotions).

size and color of the bubble representing an event is based on the importance of
the event which is represented by its climax score. Again, hovering the mouse
cursor over an event shows more information. The second time-based plot is a
scatter plot that displays the climax score of individual events. Events from the
same group are represented by the same icon. In this graph, the user can select
events by dragging the mouse cursor over them.

The third view addresses the provenance of the data visualized in the other
two views. It displays a table that contains the "raw" data, allowing the user to
cross-reference events, participants, and groups with the original sources.

To facilitate visual exploration of the data, graphs in Storyteller are dynamic
and interactive. The user can filter or select data by clicking components of the
graphs or by dragging the mouse cursor. Filters activated in one view are dynam-
ically applied to the other views. Filters in one view are applied in succession,
each time reducing the amount of data on the screen. This allows the user to
drill-down into the data set, gaining knowledge of its composition in the process.
Every filter that is applied to the data can be removed independently from the
other filters.

Storyteller is open source software, built to be as generic and reusable as pos-
sible.[5] It displays data in a JSON format, specified in the documentation. This
open outfit allows anyone to visualize their data, and to modify the application
to their needs.

4   Visualizing the Data

Storyteller was used to visualize data sets from three different projects. In this
section, we discuss the projects, the data sets, the data set-specific mappings
to Storyteller concepts (i.e., *events*, *participants*, *groups*, *climax score*), and how
the visualization contributes to stimulating insight into the data. The three
projects are NewsReader Sect. 4.1, BiographyNet Sect. 4.2, and Embodied Emo-
tions Sect. 4.3. In Sect. 4.4, we summarize the differences between the three data
sets.

4.1   Newsreader

NewsReader aims to identify *what* happened to *whom*, *when* and *where* in large
amounts of news [13]. NewsReader developed NLP pipelines that identify events,
their participants, time expressions and relations between them in English, Span-
ish, Italian and Dutch text. Using this extracted information, documents which
mention the same event can be identified and a single representation of this
event can be provided. This leads to deduplication of redundant information
and aggregation of information in cases where documents provide complemen-
tary information on the event. Each event representation is linked to the places
where it was mentioned in text. This link allows comparing the different per-
spectives various sources provided on a specific event.

---

[5] https://github.com/NLeSC/UncertaintyVisualization.

The data is represented in RDF, using the Simple Event Model ([12], SEM) to model events, their participants, location and time. The extracted events form so-called Event-Centric Knowledge Graphs [9]. These graphs provide the means to investigate, for instance, who was involved in a specific event, what else did this person or company do, and who does this person interact with, etc. These are exactly the kinds of stories Storyteller was designed to visualize, using both actor-centric and event-centric representations.

The visualization we implemented displays the outcome of this pipeline. The *events* in the NewsReader Storyteller are event instances plotted on the lines of the participants involved. *Participants* are entities, e.g., a person such as Steve Jobs, or a company like Apple Inc. If two participants took part in the same event, their lines cross at the event intersection. Figure 2 shows that, according to news articles from 2003 to 2009, Apple Inc. participated in events together with, for example, the EU, Steve Jobs, and Wallmart. In this context, the co-participation graph allows the user to discover patterns in interaction: who is interacting with whom? Who else are they interacting with?
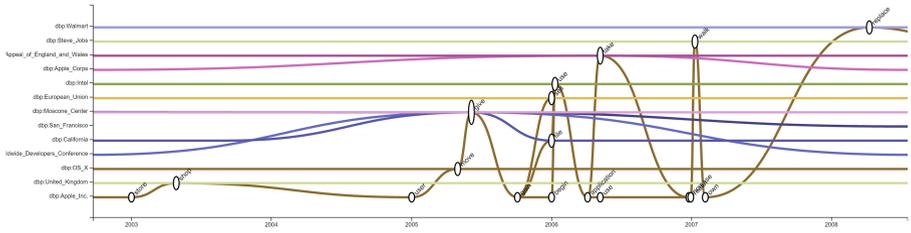
Fig. 2. Apple Inc.'s co-participants in news articles from 2003 to 2009.

Figure 3 shows the second view of the NewsReader visualization. It displays events that are connected in storylines: each event that is part of the same story is represented in the same row. Events that are mentioned in a topical relation and involve the same participants are considered to be part of a story. In Story-teller concepts, a *group* is a story. The groups are ordered based on the highest occurring *climax score* in the story. This score is based on the number of times an event is mentioned and the place in the article where it is mentioned. As defined in [14], events with the highest climax score represent critical turning points around which stories unfold. Preceding events lead up to the climax and follow-ing events resolve it. In Vossen et al. [14], stories are constructed by connecting events to climax events using bridging relations. The view below the storyline view places all events on the Y-axis based on their climax score. The color and shape of the events indicate the story the event belongs to. This visualization provides insight into how stories work up to their climax.

When taking the different levels of visualization together, users can detect patterns in the behavior of certain participants by looking at interactions, the individual events and stories over longer stretches of time. For instance, does a
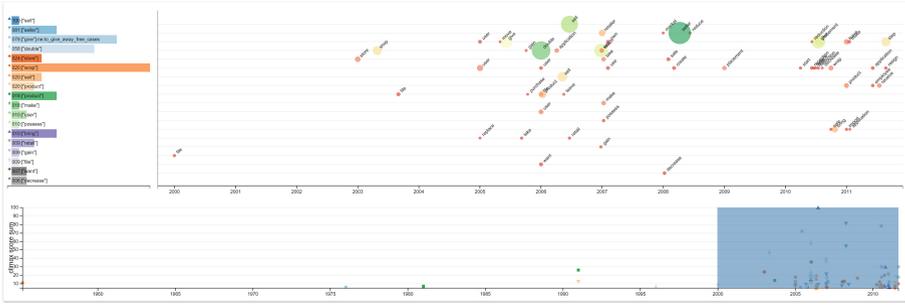
Fig. 3. Storylines containing participant Apple Inc.

specific company tend to have long-term interactions with business partners or do their dealings tend to remain limited to few events in a short period of time? What happens to companies when a specific CEO joins the company? Do they recover from a crisis?

4.2 BiographyNet

The BiographyNet project enriches the metadata from the Biography Portal of the Netherlands,[6] by running the full texts of circa 140,000 short biographies through an NLP pipeline. Previously missing structured data on topics like parents and education are automatically mined from the text, providing new leads and computerized analyses. A visualization tailored to BiographyNet's specific needs is now under development, but the BiographyNet data also serves as a good use case here to see if Storyteller can cope with this data as well.

For BiographyNet, the *participant* in Storyteller is a person, usually the biographee, about whom information is provided. The *event* is a particular thing that happened to this individual at a certain point (or period) in time. So if BiographyNet states that the Humanist philosopher Erasmus was born in 1469, the participant is Erasmus and the event is 'born in 1469'. The *group* this event belongs to is 'dates of birth'. The *climax score* is the overall number of mentions of a certain event. For this data, that boils down to the number of people from a group that were involved in the same kind of event in the same year. If in one group of people the event 'born in The Hague' occurs frequently, it will get a high climax score because it is the event that binds most people. This will allow for an easy recognition of patterns: if in a group of 300 women from the period 1950–2000 the events 'born in Amsterdam' and 'studied in Leiden' get high climax scores, than we have singled out a potentially important pattern, a distinguishing feature, for this group, which we can easily compare to the climax scores for any other groups.

The data set currently visualized in the BiographyNet Storyteller consists of important events in the lives of 25 women, 24 of whom died between 1940

---

and 1945, and one who died in 1949.[7] In addition to dates of birth, death and marriage, we visualize events that can tell us something about their 'claim to fame': 'publish', 'organize', 'found', 'paint', 'study', etc. Figure 4 shows the co-participation graph of this data set. In accordance with how the data was selected, the event(s) most co-participated in is 'death'. Figure 5 shows the second view. The large green circles in the scatter plot show again that 'death' is the most important event. The current visualization only maps events to participants and time, but other aspects such as location can easily be added.
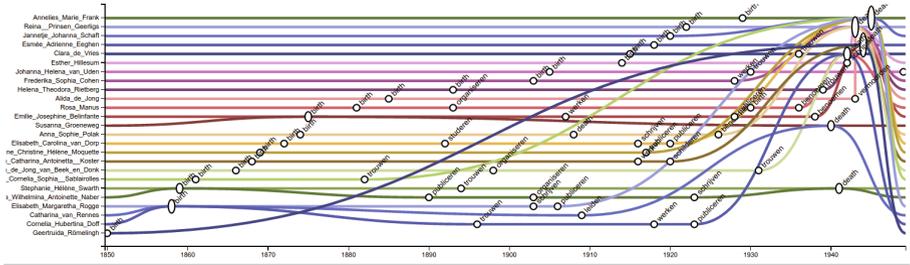
Fig. 4. The co-participation graph of the BiographyNet Storyteller data set.
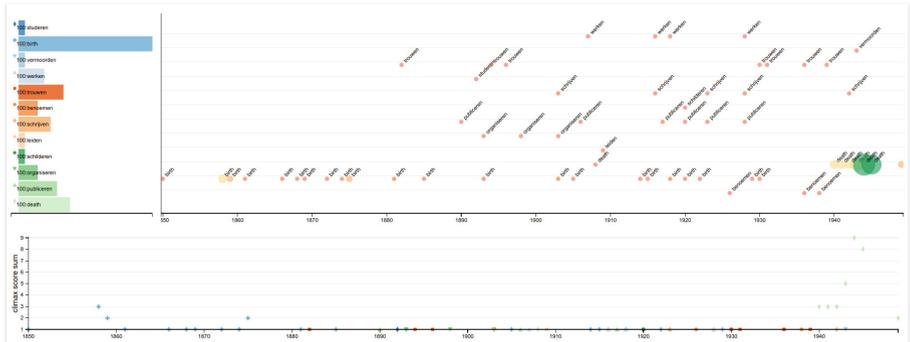
Fig. 5. The BiographyNet Storyteller data set in the second view.

For future work it will be a challenge to visualize uncertainty when the year in which an event happened is not known or uncertain. For example, Eramsus was born circa 1469, but the exact year is unknown. For other cases, a date or year could be missing completely and sometimes there is a period in which something happened – e.g., a person studies Law for several years; a person had a daughter at the end of the eighteenth century (see also Stokes (2015) [11]).

---

[7] This last woman was added for reasons of visibility: Storyteller's timeline ends at the last event and those of 1945 were hard to see without anything occurring afterwards.

4.3   Embodied Emotions

The goal of the Embodied Emotions project was to investigate the relationship between emotional expressions and body parts in Early Modern Dutch theater texts [6,15]. The main focus was on tracking changes in the embodiment of emotional expressions over time. To this end, a Historical Embodied Emotions Model (HEEM), was developed. The model consists of two layers: concept types and emotion labels. The concept types, i.e., Emotion, Body part, and Bodily response, refer to the embodiment of emotions. In addition, there are 38 historically accurate emotion labels. A representative set of 29 Dutch theater texts from 1600–1830 was manually annotated with HEEM labels. The manual annotations were used to train text classifiers [15], which were subsequently applied to a bigger set of 279 theater texts [16]. Body parts were further specified in an additional data processing step. First, all words marked with HEEM label 'Body part' were manually assigned a more specific body part label, e.g., Heart, or Head. The resulting mapping was applied to the bigger corpus; for each sentence that was assigned HEEM label 'Body part' and that contained words from the body part mapping, the corresponding body part label was assigned to the sentence as well.

A data set for Storyteller was generated from the manual annotations. This data consists of various, interrelated levels that could be visualized:

1. The expression of emotions in a singular theater text
2. The expression of emotions over time over a group of texts (e.g., divided by genre: comedy, tragedy, farce)
3. The bodily location of emotions in a singular theater text
4. The bodily location of emotions over time over a group of texts
5. The emotional qualities of body parts over time
6. Provenance of all the projections

For the demo, we decided to focus on visualizing the relationship between emotions and body parts over time. In Storyteller concepts, *event*s are used to refer to an emotion expressed in a single theater text (e.g., Love in Vondel's *Lucifer*). A *participant* is a body part label (e.g., Heart). The *group*s are abstracted emotions (i.e., HEEM labels). The *climax score* was calculated by dividing the number of times an emotion occurs in a text by the number of sentences in the text. The *date* of a text is its year of publication.

The resulting visualization provides an environment where researchers can switch between events (emotions) and participants (body parts), and represent their interrelations over time. The co-participation graph enables researchers to explore what emotions are felt in which body parts, and what body parts are used to express specific emotions. Figure 6 shows the co-participation graph of Heart, which is the body part occurs most frequently in the data set. Another result that was found by interacting with the visualization is that anger can be expressed through the eyes, the tongue, the hands and the nerves. The time-based graphs in the second view of the visualization enables researchers to gather evidence for statements such as "the heart has become the most important seat

Fig. 6. Co-participation graph of the Embodied Emotions data filtered on Heart.

Fig. 7. The second view of the Embodied Emotions data filtered on Heart.

for the expression of love since 1750", and "body parts like finger nails and hair were very emotional around 1600, but are 'neutralized' after 1700". Figure 7 shows the second view with the Embodied Emotions data set filtered on Heart. The column-like distribution of points in the graph indicate that the data set contains texts from certain years only. By filtering the data researchers can take into account domain knowledge about general developments over time, e.g., the fact that theater texts in general become less explicitly emotional around 1800. Filtering also allows researchers to focus on a "basic" set of emotions, by removing more "moral reflective" emotions, such as honor, shame, and feelings of guilt.

4.4   Summary

The three projects we described all have different data and use cases: direct interaction of participants through modeling specific events and their stories (NewsReader), similarities between people by representing the kind of events they are involved in and when (BiographyNet), and the historic development of how people related emotions and bodyparts (Embodied Emotions).

Table 1. Events, participants, groups, and the climax score in the three projects.

Concept	NewsReader	BiographyNet	Embodied Emotions
Event	Event (something happening in the real or an assumed world)	Event type occurring in a specific year (e.g. marriage in 1910, death in 1943)	Emotion expressed in a single theater text (e.g., Love in Vondel's *Lucifer*)
Participant	Entity (e.g., Apple Inc., Steve Jobs)	A person	Body part (e.g., Heart)
Group	Story linking events and participants	Event type (e.g. marriage, death)	Abstracted emotion (e.g., Love)
Climax score	Prominence of event mentions (number and location in articles)	Number of people involved in this type of event in a given year	Number of times an emotion is expressed divided by the number of sentences in the text

Table 1 lists the data set-specific mapping to events, participants, groups and climax score in each of the three data sets.

5   Conclusions

Digital Humanities projects often aim at bringing order to large amounts of (semi-)structured texts, or heterogeneous data. In this context, visualizations are essential to provide humanities scholars with insight and to inspire future research. This paper presented Storyteller, a visualization tool that does just that. The applicability of the tool was demonstrated by discussing preliminary results from visualizing data from three different projects. One of the strengths of Storyteller is that *participants*, *events*, *groups* and *climax scores* can be used to refer to different things in different data sets, depending on the needs of the scholars who will use it. Access to the original data is provided to enable the user to check the findings and closely read the original sources whenever necessary.

Storyteller's first two views reveal patterns that are hard to detect by close reading. The tool visualizes complex and heterogeneous data, which is by nature hard to understand. Using Storyteller also requires a proper understanding of how the tool works. We therefore aim to develop tutorials that help scholars to understand an use the tool. Generally, the humanities scholars that explored data using Storyteller were positive about its usability. For future work, a more formal usability study is required.

The third view provides insight into where information comes from by presenting the source text, but there are many more aspects to uncertainty and perspectives that are not yet addressed. The main focus of our future work will be on visualization of uncertainty, to cope with different perspectives, uncertain time periods and contested data.

References

1. Van den Akker, C., Legêne, S.: Digital hermeneutics: Agora and the online understanding of cultural heritage. In: Proceedings of the 3rd International Web Science Conference, pp. 1–7 (2011)
2. Bod, R.: Who's afraid of patterns? The particular versus the universal and the meaning of humanities 3.0. BMGN-Low Countries Hist. Rev. **128**(4), 171–180 (2013)
3. Drucker, J.: Humanities approaches to graphical display. Digit. Humanit. Q. **5**(1), 1–21 (2011)
4. Drucker, J.: Graphical approaches to the digital humanities. In: A New Companion to Digital Humanities, pp. 238–251 (2016)
5. Jessop, M.: Digital visualization as a scholarly activity. Literary Linguist. Comput. **23**(3), 281–293 (2008)
6. Leemans, I., van der Zwaan, J.M., Maks, I., Kuijpers, E., Steenbergh, K.: Mining Embodied Emotions: a Comparative Analysis of Sentiment and Emotion in Dutch Texts, pp. 1600–1800 (To appear)
7. Novak, J., Micheel, I., Wieneke, L., Dring, M., Melenhorst, M., Garcia Moron, J., Pasini, C., Tagliasacchi, M., Fraternali, P.: HistoGraph–a visualization tool for collaborative analysis of networks from historical social multimedia collections. In: 2014 18th International Conference on Information Visualisation (IV), pp. 241–250 (2014)
8. Ockeloen, N., Fokkens, A., ter Braake, S., Vossen, P., de Boer, V., Schreiber, G., Legêne, S.: BiographyNet: managing provenance at multiple levels and from different perspectives. In: Proceedings of the 3rd International Conference on Linked Science, vol. 1116, pp. 59–71 (2013)
9. Rospocher, M., van Erp, M., Vossen, P., Fokkens, A., Aldabe, I., Rigau, G., Soroa, A., Ploeger, T., Bogaard, T.: Building event-centric knowledge graphs from news. J. Web Seman. **37–38**, 132–151 (2016)
10. Russo, I., Caselli, T., Monachini, M.: Extracting and visualising biographical events from wikipedia. In: Proceedings of the First Conference on Biographical Data in a Digital World 2015, pp. 111–115 (2015)
11. Stokes, P.A.: The problem of digital dating: a model for uncertainty in medieval documents. In: DH2015 Abstracts (2015)
12. Van Hage, W.R., Malaisé, V., Segers, R., Hollink, L., Schreiber, G.: Design and use of the simple event model (sem). Web Semant. Sci. Serv. Agent. World Wide Web **9**(2), 128–136 (2011)
13. Vossen, P., Agerri, R., Aldabe, I., Cybulska, A., van Erp, M., Fokkens, A., Laparra, E., Minard, A.L., Aprosio, A.P., Rigau, G., Rospocher, M., Segers, R.: Newsreader: how semantic web helps natural language processing helps semantic web. Special Issue Knowledge Based Systems. Elsevier (to appear)
14. Vossen, P., Caselli, T., Kontzopoulou, Y.: Storylines for structuring massive streams of news. In: Proceedings of the 1st Workshop on Computing News Story-Lines (CNewS 2015) at the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2015), Bejing, China (2015)

15. van der Zwaan, J.M., Leemans, I., Kuijpers, E., Maks, I.: HEEM, a complex model for mining emotions in historical text. In: 2015 IEEE 11th International Conference on e-Science (e-Science), pp. 22–30 (2015)
16. van der Zwaan, J.M., Maks, I., Kuijpers, E., Leemans, I., Steenbergh, K., Roodenburg, H.: Historic Embodied Emotions Model (HEEM) dataset (2016). http://dx.doi.org/10.5281/zenodo.47751