Exporatory Political Search (ExPoSe)

University of Amsterdam
Dispectu BV
Koninklijke Bibliotheek
Meertens Instituut
Nationaal Archief
Spinque BV
Tweede Kamer

CLICK//NL Top Sector
NWO Creative Industry Call 2012

Abstract

The Parliamentary Proceedings of a country provide a backbone of its textual cultural heritage. It covers centuries, in a standardized format, with highly controlled quality. Its completeness as a longitudinal corpus and its content covering a large part of a nations history, makes the proceedings both of interest in their own right, and as a tool to contextualize other data collections. The Dutch Parliamentary Proceedings are available in digitized format starting from 1576 (although large parts of 1625-1815 have not been digitized). However, access is poor as it is limited to full text search of OCR'ed text editions, or metadata descriptions of the archives. The main issues and the *main research problems* of the proposal are: How to turn the corpus of digitized heritage texts into a connected network of information? How to exploit the obtained networked structure for interactive exploratory search?

The ExPoSe project aims to bring out the full potential of the Dutch Proceedings by transforming them into a structured XML format. To really function as a backbone, it must be easy to connect other documents deeply into the proceedings. We make this possible by connecting the content of the Proceedings to the Linked Open Data cloud, which we extend with several historically relevant sources (among them the huge newspaper collection of the KB and the *Woordenboek der Nederlands Taal*, WNT). The project will run for 5 years and employ 2 PhD students and one .5 FTE programmer. The *consortium* contains the complete chain from raw data producer to end-users in the Creative Industry. It consists of 5 public and 2 private partners, contributing 315K and 83K Euro, respectively.

1 Description of proposed research

The motivation of the ExPoSe project is based on the simple observation that:

1. Research is done by establishing and following connections between pieces of information.
2. Much new technology consists of simplifying the establishment of connections (e.g., URL on the web, WikiWord in Wikipedia, Hashtag on Twitter). The Linked Open Data (LOD) Cloud is meant to serve as a backbone of this web of interlinked information.

These two points explain the remarkable success of the web: the ease by which information can be linked on it has caused an enormous increase of the value of that same information. In addition, we observe that:

3. Explicit connections between pieces of information are almost completely lacking in the newly established digitized cultural heritage collections.

This hinders bringing out their full potential both for research, for society, and for the economy. Arguably this is the single most important problem facing digital cultural heritage collections. Of course the digitized texts are full of *implicit* connections: they may discuss the same events, mention the same entities, are about the same topic or concept, etc. This leads to our main research problem:

Research Problem How to to turn the corpus of digitized heritage texts into a connected network of information. What can serve as the backbone of this cultural heritage web? How to turn implicit links into explicit links? How to adapt modern tools for analyzing the WWW to the cultural heritage web?

If we are able to make these implicit connections explicit, the digitized collections gain in value similar to the WWW.

Our approach to this is the following. We create a backbone (spine) for this cultural heritage web consisting of two main parts. We observe that many important events, entities and topics in Dutch history are also discussed in the parliament of the Netherlands, and thus they occur in the parliamentary proceedings. Viewed in this way, the parliamentary proceedings are a complete longitudinal corpus with regular (weekly) measurement points, consistent data model, and very high quality. This leads to our first key objective:

Key Objective 1 (Data). To bring all parliamentary proceedings together, in a common rich format that captures the internal structure of the debate, as well as provides semantic annotation on the speakers, their role and political affiliations, the topic of the debate, etc.

Hence one can view the parliamentary proceedings as a historically oriented counterpart of the LOD, provided that events, entities and topics are identified explicitly, and linked to contextual information in resources like Wikipedia/DBpedia, the *Biografisch Portaal*, the newspaper archive of the KB, the dictionary of Dutch Language (WNT), etc. This leads to our second key objective:

Key Objective 2 (Links). To link historic events, entities and concepts into a network which integrates all these datasources, using the timeline of the parliamentary proceedings as the backbone.

The resulting collection will have unprecedented power, with a wealth of information in semantic annotation and links within and between documents. This creates a need for tools that help searcher explore this rich content without the need to acquaint themselves with intimate details of the encoding schemes of the data. This leads to our third key objective:

Key Objective 3 (Tools). To develop powerful tools that allow searcher to explore the rich content, by interactively constructing complex queries or search strategies, and interactively exploring the results of each stage.

This envisage solution has many desirable consequences:

- Network of interlinked Dutch heritage documents with the Parliamentary proceedings at its core and based on a bipartite graph linking events, entities and concepts in texts to their representations in the historic LOD.

- For Humanities scholars understandable and usable tools for recognizing events, entities and concepts in digitized texts, and mapping these to the historic LOD. These tools are based on existing named entity and event taggers and text classification technology.
- It is important that the tools are transparent and tunable by users. To achieve this the map will not be a binary one-to-one mapping but a probabilistic one to many mapping (from each text we create an event, an entity and a concept model in the form of a language model).
- A data sharing, linking and provenance model that keeps data-curators happy. In particular, they still get acknowledged when their data is used (by unique users, clicks, number of downloads, etc).

1.a Innovation network and crossovers to other innovation networks

The ExPoSe project has great potential impact on the research agendas as laid down by the Cultural Heritage and Media and ICT innovation networks.

To the *Cultural Heritage* network, the proposed project makes clear contributions to all current trends of the innovation agenda:

- *Trend 1: Connected Heritage*. Here, the ExPoSe contribute the central problem of making implicit connections explicit, through semantic annotation and linking of textual corpora. The historic LOD as envisaged in ExPoSe goes far beyond the ideas of the innovation agenda, by using the parliamentary proceedings as a longitudinal backbone covering main events in Dutch and global history, while still linking to a range of other relevant resources (such as Wikipedia/DBpedia).
- *Trend 2: New Users*. The proposed project is very user-centered, working on use cases from data journalists, historians, archivists, and the general public, and developing tools tailored to their needs. The design to connect resources from various institutes, rather than integrate them in a central service, will keep the role of the original curators and custodians and make traffic and revenues visible and sharable.
- *Trend 3: New Multidisciplinary Research Methods*. The development of tools that allow for complex search strategies each corresponding to a novel digital research method is key to the setup of ExPoSe. Complex (re)search strategies van be saved and shared, and adapt to new cases. We envision that tools for data journalists will also be valuable for historical research, and vice versa, introducing novel perspectives for both groups of users.
- *Trend 4: Technology as a Mediator*. Adequate tools are crucial in unleashing the power of richly structured corpora, yet this runs the risk of reducing meaningful information interaction to the consultation of 1 or a few resources. The tools of ExPoSe are designed to be controlled and operated but the searcher who actively explores the available data, and actively let them explore (significant slices) of the collection, rather than just the top ranked handful of results.

To the *Media and ICT* network, the main contributions are:

- With respect to the theme of Cultural Heritage, the ExPoSe project is central to the suggested problems: interoperability; enriching collections; contextualization; rich interaction and visualization; etc.
- The proposed project has clear connections with the trends in data: mass digitization, open data, digital curation, etc.
- There are also clear links relations the trends in technology: new services, advanced search, cloud computing, open source, etc.
- Finally, the is a clear connection with the trends in users: user empowerment, time-aware and location-aware search, etc.

The ExPoSe project has both media (data journalists, traditional and online newspapers) and ICT (advanced tools for annotation, advanced search, novel exploratory search tools) as central themes. The contributions to *Cultural Heritage* and to *Media and ICT* may come as no surprise given the applicants of ExPoSe were active contributors to both these two innovation networks, and the proposed project provides a natural cross-over of cultural heritage and media and ICT innovation agendas.

There are more modest contributions to the other innovation agendas:

- The *Smart Design* agenda highlights co-design and co-creation, which is central to the ExPoSe living laboratory.
- The *CI NeXt* agenda highlights novel forms of organization: crowdsourcing, open data, open the archive, and apps for democracy, which are used by ExPoSefor realizing its objectives.

We use these approaches, in areas where this is still rare, and will deliver key insights in the application in these domains (use case, best practices). This is of great practical importance, but we do not envision significant contributions—our main research questions are in different areas.

1.b Scientific quality

Key objective 1 (data) is to bring together the whole parliamentary history in a common format and make the internal structure of the debates machine readable. This will address several concrete research questions:

- Parliamentary archives are curated by the National Archives, yet scattered over various different collections, including: *het archief van de Staten-Generaal, 1576-1796* (1.01.02); *de archieven van de Wetgevende Colleges van de Bataafse Republiek en van het Koninkrijk Holland, 1796-1810* (2.01.01.01); *de Handelingen van de Eerste en Tweede Kamer der Staten-Generaal over de jaren, 1814-2009* (2.02.21.01). Remaining archives (1810-1814) are in the *Archives de l'empire* in France, as the Netherlands was part of the French Republic in those years.

  Can we bring together all these sources and provide access to the whole parliamentary history?

- Parliamentary proceeding 1815–1995 are digitized and made available at http://statengeneraaldigitaal.nl/ (KB, Tweede Kamer), and proceedings (1995–now) are published at http://www.officielebekendmakingen.nl/ (http://overheid.nl/). An integrated collection of all proceedings 1815-now, in an enriched XML format, is already available from the preceding project politicalmashup.nl.

  Earlier parliamentary proceedings are not completely digitized, the national archives host incidental scans and transcriptions, and ING/Huygens offers a digitized version of an earlier printed edition of the proceedings of 1576–1625 at http://www.historici.nl/retroboeken/statengeneraal/. The budget does not allow for large scale digitization and transcription, but we have the resources to do pilot experiment in collaboration with consortium partners such as the National Archives.

  Can we bring together all existing digital transcripts in a common format?

  Can we rely on contextual archival descriptions and further finding aids to provide access to the remaining content?

  Can we mix such metadata descriptions and transcripts together to improve access of each individually?

Key objective 2 (links) is to link the parliamentary proceedings to a wealth of other resources, creating a historic LOD. This will address a range of concrete research questions:

- What is the coverage of the current historic LOD cloud (Wikipedia, *Biographisch Portaal*, KB *krantenarchief* and WNT) for historic events, entities and concepts occurring in the Parliamentary proceedings?
- How useful (complete, correct, standardized or standardizable) is the metadata common in this LOD cloud? How does it decline over the years?

- What are the most effective (in terms of annotation costs and benefits to the researcher) adjustments to text extraction and annotation tools for modern texts in order to achieve useful performance on historic texts?
- With what quality (precision/recall) can we use the entity, event and concept models of documents to link these documents to non-textual media with sparse descriptions (like *BeeldBank* photographs, film/video from *Beeld and Geluid*, portraits and paintings)?
- What tools can help the researcher (historian or data journalist) with the transition from investigating very few high quality impressionistic data points to representative, large scale but noisy and probabilistic datasets?
- Can the OAI-PMH protocol with digital object identifiers and resolvers and XML encoded documents perform with the same success as the HTTP, URL, HTML triplet does for the web?

Key objective 3 (tools) will build a range of tools for exploratory searching the political history and resulting historic LOD. This will address a range of concrete research questions:

- Standard search tools are tailored to the *topical relevance* of the content, focusing exclusively on the "what" is said. Can we build tools that focus not only what is said, but also by who and to whom, and why?
- Modern Web formats has rich semantic annotations, yet unleashing these powerful cues required mastering a complex querying language, and significant "programming" skills. Can modern insights in exploratory search on structured data bring this power into the hands of researchers and the general public?
- The resulting tools exploit the semantic annotations to bring out what remains hidden in the plain text: the actual political process and strategies within a debate, as well as how the politics evolved over time – of politicians, parties, and the political system as a whole. Can this provide a new data-driven perspective on both current politics and our political history?

Data Driven Historical Research We claim that *exploratory search* is the best way to bring out different aspects of complex textual data collections. The same tools can be used by scientists on historic material and data-journalists on current material. We make this concrete with a few examples, using pilot prototypes developed by our private partners on the proceedings data from 1814–2012.

Figure 1 shows a prototype application (build by Spinque) for exploratory search over the semantically annotated parliamentary proceedings. Searchers can start in any way, e.g., search for a given topic, and then zoom in on particular speakers, parties, dates, etc. In this example, our searcher starts with an initial query "*kernwapens*" (nuclear missiles). The right side pane displays various aspects of the whole result set (potentially contain millions of speeches) rather than only the top results, in simple information graphics, inviting searchers to explore various different facets.

Figure 2 shows a refined query, zooming in on a particular party (right wing "VVD") and on opening statements of a member of parliament addressing the house (rather than an interruption to another speaker). Although this is a trivial example, it illustrates the way in which searchers can interactively construct a complex query without knowledge of complex query languages or the underlying encoding schemes.

The side bar's facets are aggregated over large sets of data. The resulting simple information graphics (how much relevance can be attributed to facet X) are ideal for exploring the data and various slices of the data. Such plots may use any pair of variable (e.g., counts over parties or speakers, relative to a topic), may use various ways of aggregating spreaders (e.g., parties, but also based on external data, such as biographic data showing age, gender, education level, height, etc., of the members of parliament). One could even go far beyond this, for example, by showing how occurrences vary over time, for each speaker, party, or other aggregate.

Figure 3 shows a prototype ngram of phrase viewer (build by Dispectu) , able to trace the frequency of phrases uttered in the parliament. E.g., in this case the phrase "*cultureel erfgoed*" (cultural heritage) is traced over time, showing how this terminology emerged only in the 1980s. This highlights the temporal

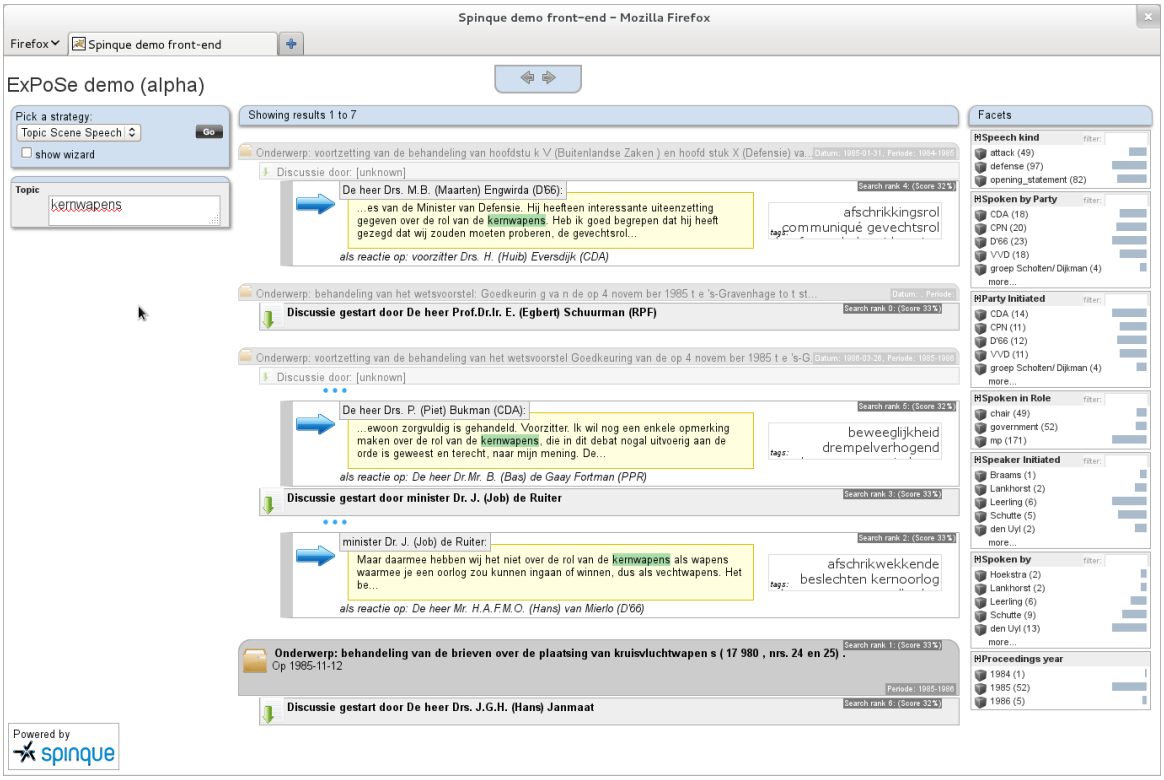

Figure 1: Exploratory search (prototype): search for a topic (*kernwapens*: nuclear missiles) on the level of individual contributions, right hand pane lists various frequencies of speeches per speaker, party.

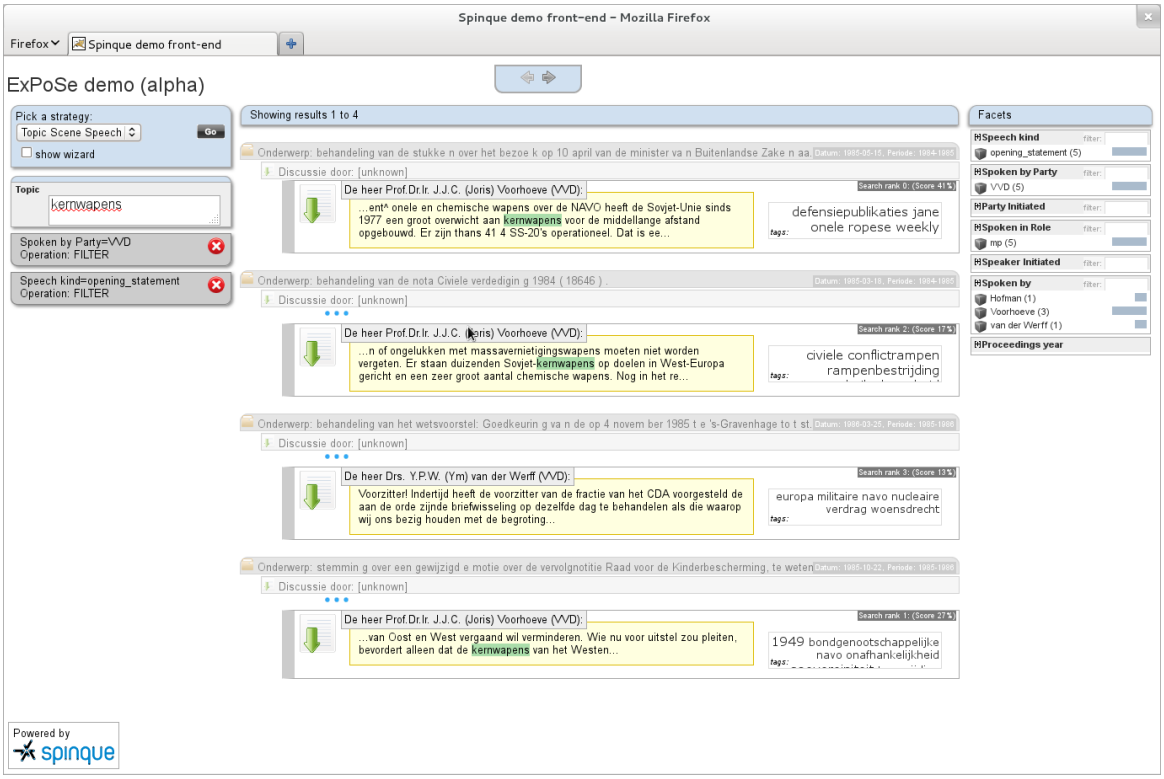

Figure 2: Exploratory search (prototype): query *kernwapens* (nuclear missiles) refined by party "VVD"; refined by opening statements (first utterance of a speech to the house, not an interruption.

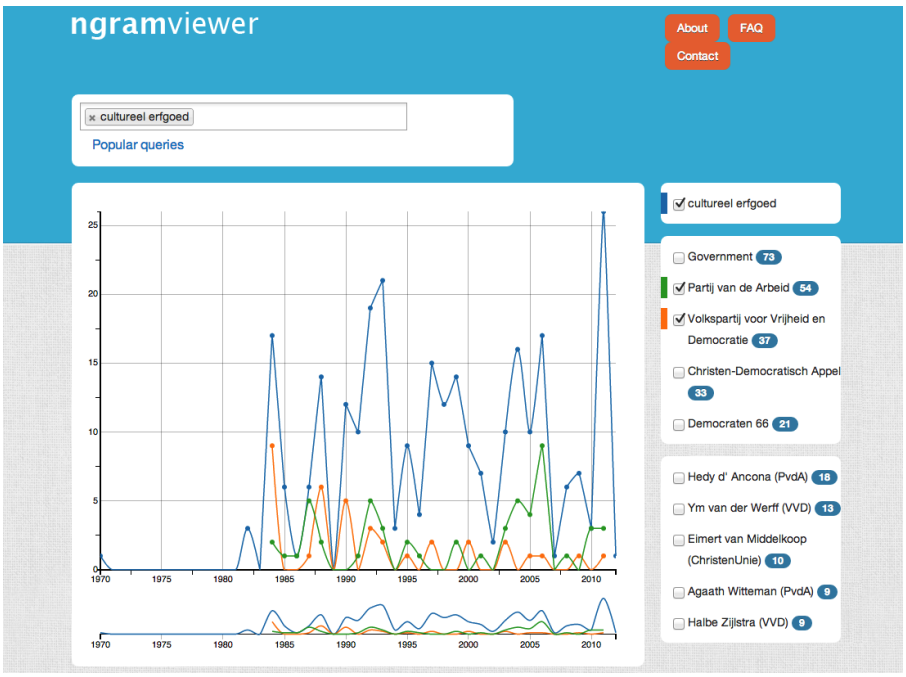

Figure 3: Frequency of n-gram occurrence over time (prototype) showing "cultureel erfgoed" (cultural heritage) as a phrase occurring in the proceedings: overall data plus occurrences in speeches of right-wing VVD party and left wind PvdA party.

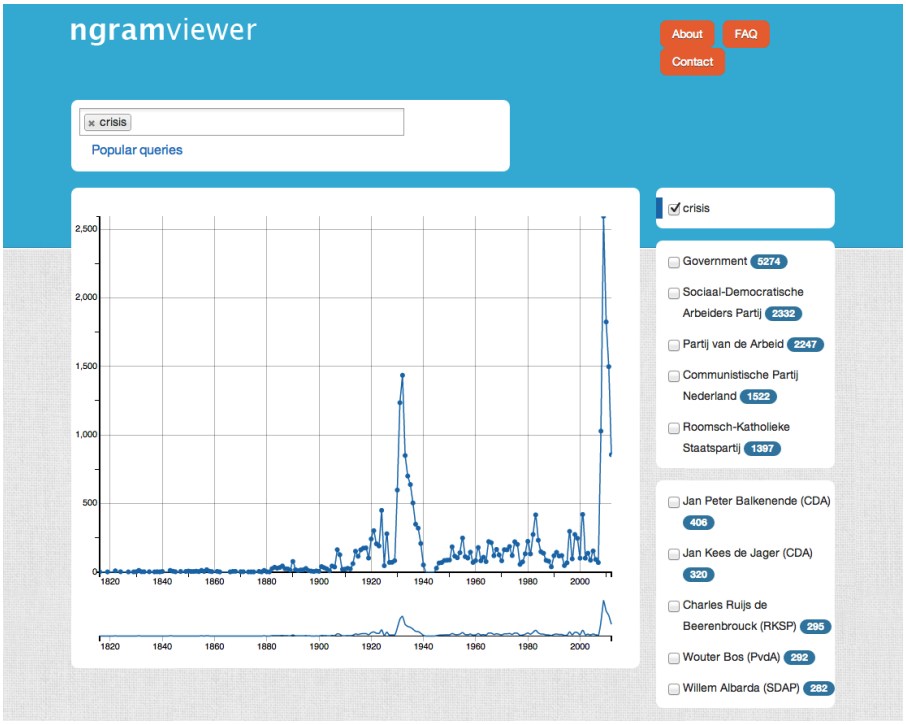

Figure 4: Frequency of n-gram occurrence over time (prototype) showing the frequency of the word "crisis" being said in parliament.

facet. Similar to the exploratory search prototype, a user can break down the utterances over the respective speakers and their parties. The top 5 political actors with the relative frequencies are indicated on the right hand pane. The figure shows a comparison between the right wing party "VVD" (downward trend in recent years) and the left wing party "PVDA" (upward trend in recent years). Such analysis facilitates insight into agenda setting (which party puts a topic on the agenda first, and which parties follow?) and the use of framing (the use of certain words to influence public opinion; "taxpayer money" versus "government funds") by political parties.

Figure 4 shows the number of occurrence of the word "crisis" in the Dutch parliament. There are very

clear peaks during the great depression (1930s) and in the more recent years (global European recession). This gives insight in the state of the world, similar to the Google Zeitgeist, Yahoo Buzz, and r-word index of The Economists (an indicator of economic activity based on occurrences of the word "recession" in the New York Times and Washington Post). Links between the Proceedings and the KB newspaper corpus can be used to *explain* and further explore peaks and trends.

1.c Utilisation/relevance

The ExPoSe project has as partners the most directly affected institutions in term of the data (Dutch Parliament, National Archives of the Netherlands, National Library of the Netherlands), in terms of tools and technology (Spinque, Dispectu, Meertens/Nederlab), and in terms of representatives of the prospective user groups (data journalist, e-humanities scholars, historians, archivists). These partners will together form the project's living lab, and help shape the project's results as active co-creators providing crucial feedback and ideas in all stages of the project.

Leading Dutch organizations for promoting democracy and citizen participation are already using data and tools supplied by PoliticalMashup. Its Parliamentary Proceedings 1814-2012 set formed the largest dataset available at the September 2012 "*Apps for Democracy*" hackaton held in the Dutch Parliament. These organizations include *KiesKompas BV*, *Stichting Het Nieuwe Stemmen*, *Stichting Hack de Overheid*, Open State Foundation, and *Stichting Netwerk Democratie*.

A good example of knowledge transfer and exploitation of data is nupubliek.nl, part of the leading Dutch online news site nu.nl owned by Sanoma BV. Using a daily crawl of the PoliticalMashup database of proceedings, parties and politicians, nupubliek.nl semi-automatically creates profiles of politicians, and attaches politicians and proceedings to related news articles.

More generally, with the massive increase in information available nowadays, there is a increasing need for tools that help us find the right information at the right time, to make the right decision. Paradoxically, despite the volume of information available, searchers on the Web are known to seldomly look beyond the fold (the number of results visible without scrolling down or clicking next). Each query generates millions of results, yet the number of clicks below the the tenth document is negligible. This creates a need for novel information access technology that invites us to look deeper, actively explore and engage with the content, by supporting searchers to articulate complex queries or search strategies, and explore the whole result set (rather than a handful of top ranked results) at every step in this process [1].

The project will set-up a living laboratory for searching political history, which will greatly facilitate the project, but also has profound potential effects on the researchers and journalists involved in these labs. As stated above, the tools invite searchers to articulate complex search strategies. These strategies can be stored and shared, allowing users to learn from each others and holding the promise to lead to collaborative research that goes far beyond what's possible now by an individual researcher.

Let us consider the recent data, which is of great interest to journalists and the public (civic journalists). The ExPoSe tools enable them to explore the data in far greater depth, yet require no special training. We will actively collaborate with the data journalism user group to develop tools that are tailored to their needs, with the ambition to develop generic tools that will allow them to develop their own research tools formulated as complex strategies and visualizations of exported data. Let us consider the historic data, which is of great interest to the historians, archivists, and the general public. Also here the ExPoSe tools will enable them to explore the data in unprecedented ways. What now if we apply the tools of the data journalists to the historical data? And the digital research methods of historical research to recent data? This will allow for new cross-fertilization between the two groups, and turn search into research, and research into search.

Moreover, the resulting combined data sets can be used in unexpected ways. The whole parliamentary history is a longitudinal corpus covering almost 500 years, making tools like the ngram viewer fit to study the history and evolution of the Dutch language.

1.d Cohesion of the research

The project is greatly facilitated by the availability of basic versions of all the needed components.

This leads to three work-packages:

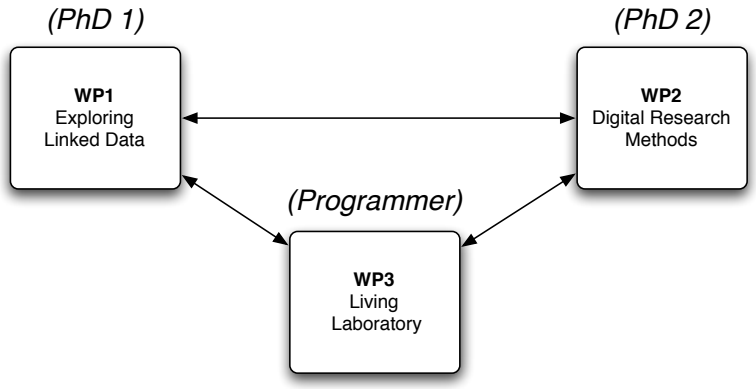

Figure 5: Work-packages and relations.

WP1: Exploring Linked Data In this work-package, we will develop a range of tools for exploring linked data, with the parliamentary proceedings as its core (corresponding to Key Objective 2). This includes tools for enriching the data by further semantic annotations and links to other resources, and novel access tools supporting exploratory search over the linked data. The main focus is on recent data and the data journalism user group. WP1 will be the main responsibility of the *PhD Student 1*.

WP2: Digital Research Methods In this work-package, we will bring together the whole parliamentary history in a common format bringing out the internal structure of the debate, with a range of semantic annotations (corresponding to Key Objective 1). This includes search over a variety of different resources (literal transcriptions of parliamentary proceedings, contextual archival descriptions of the proceedings, indexes on the proceedings, etc). The main focus is on historic data and the historical research user group. WP2 will be the main responsibility of the *PhD Student 2*.

WP3: Living Laboratory In this work-package, we set up a living lab for political (re)search and implement access tools aiming for incremental query or search strategy construction and interactive result exploration. The search tools allow for i) expressing complex needs as search strategies, and ii) exploring the results as simple information graphics. WP3 will be the main responsibility of the *Programmer*.

Figure 5 shows the dependencies between the work-packages: WP1 builds on the infrastructure of WP3, and develop novels tools for exploring linked data, again contributed to WP3. WP2 builds on the infrastructure of WP3, and develop new digital research methods on the combined resources, and again contribute these to WP3. Conversely, the robust prototype tools of WP3 allows ranges of experiments for both experts and prospective users, allowing for a spiral development of WP1, WP2, and WP3. Finally, in the final fifth year, WP3 builds a final application and collects all project results in the knowledge transfer phase.

Risks *The work plan has many dependencies, will this cause a risk for overall project progress, if problems or delays are encountered in one of the streams?*

Our solution is to plan activities in such a way that each can progress independently. Note that in a way, each of the individual streams would be a suitable project in itself. In our experience, more and faster insights can be obtained (and even in an easier way) by approaching this complex problem from many directions, and have one community learn from the other etc. Such cross-overs lead to far greater innovation power than a traditional risk avoiding project. This more organic approach is dominant in the creative industry. We do have positive experience with this in the NWO/CATCH program where researchers are primarily based in a cultural heritage institution (the *laboratorium extra muros* concept), and in related projects with national partners, and with European partners in EU projects.

Is the project viable given the modest budget?

Yes, because our approach is to "cash in" on massive amounts of earlier work. Most notably Marx's http://politicalmashup.nl/ project and Kamps's http://staff.science.uva.nl/~kamps/readme/ project, but also consortium partners have invested in publishing the proceeding 1815–1995 (http://statengeneraaldigitaal.nl/) and in providing access to the combined archives of the Dutch Parliament. Hence we stand on the

shoulders of giants. As a personal note: we (as applicants) are honored by the massive support from so many leading institutions in the consortium.

The project has a unprecedented ambition, why not focus on a single clearly defined component?

First, it is our experience that problems get easier when daring to think beyond the obvious next step. Hence we propose an ambitious and open ended project, providing a clear vision of where we would want to be, but still proceed step by step with this clear goal in mind. Second, while it would be straightforward to turn the proposal into a large scale project involving many research teams and a long duration, we fully understand the decision to work with moderate budgets. The consequence is that we will encounter many interesting new problems along the way, and may even see potential solutions to them, but without the capacity to pursue them properly. These new problems and potential solutions may ignite follow-up projects in the Creative Industry.

Consortium Meetings Given the size of the consortium, a more formal organization of meetings will be set up. All meetings are open to all consortium members (either in person, or listening in over Skype), and notes will be made available on the consortium site.

- The PIs and researchers will meet on a weekly basis;
- Once a month a half day meeting is scheduled, in which the researchers plus relevant consortium partners come together and discuss progress, opportunities.
- Every six months a full day meeting is organized, with alternating an annual review meeting (months 12, 24, 36, 48, 60) targeting the extended consortium, and in between a theme meeting devoted to a particular aspect of the problem and project targeting also researchers and practitioners outside the consortium. These meetings will be held at varying consortium partners.

Information meetings will occur natural and frequently, partly by basing the two researchers at the Dutch House of Representatives in The Hague for one day a week, in close proximity to the National Archives and National Library. The researchers will also spent time (based on the need) at the other partners in Amsterdam.

The ExPoSe living laboratory will continuously offer additional data, links, and tools throughout the project, and actively publicize these beta services, allowing us to log traces of interaction activity and direct feedback. This will also greatly facilitate keeping the whole extended consortium up to date with all project results.

References

[1] J. Allan, W. B. Croft, A. Moffat, and M. Sanderson. Frontiers, challenges, and opportunities for information retrieval: Report from swirl 2012 the second strategic workshop on information retrieval in lorne. *SIGIR Forum*, 46(1):2–32, 2012.