Center for Digital Humanities
Research Proposal

1 Aanvragende organisatie

UvA

2 Private partner(s)

Spinque; Koninklijke Bibliotheek.

3 Titel projectaanvraag

ExPoSe: Exploratory Political Search

4 Keywords

XML Retrieval, Exploratory Search, Complexly Structured Data, Parliamentary history.

5 Projectomschrijving

Parliamentary proceedings and other transcripts of meetings are a common document genre characterized by a complex narrative structure. The essence is not only what is said, but also by who and to whom, and why. Standard search tools are tailored to the topical relevance *of the content, focusing exclusively on the "what" is said. Modern Web formats based on XML allow for semantic annotations like the speaker of each speech in order to capture this debate structure, as well as the related content of debate. Up to now, unleashing these powerful cues required mastering a complex querying language, and significant "programming" skills, yet modern insights in exploratory search on structured data hold the promise to 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 – providing a new data-driven perspective on our political history.*

The ExPoSe project brings together two related research lines that have attracted considerable interest in recent years. On the one hand, the Dutch parliamentary proceedings have been harvested and semantically enriched in the `http://politicalmashup.nl/` project. On the other hand, there are recent insights in searching structured collections (in particular those in museums, archives, and libraries) by both experts and novices suggest novel exploratory search methods tailored that are tailored to their specific demands. Generic search methods, as developed for superficial navigation of the Web, fail to support scholars properly since they cannot deal with the unique characteristics of the material, the specific demands of the researchers using the material, and context of their special use-cases. Rather than returning ten-blue-lines in response to a two-word query, the resulting system will support scholars during a whole task or search episode, by iteratively constructing complex queries, and interactively exploring the whole result space at every stage.

The aim is to build a public search system that exploits our project results for a large archive of Dutch political data. The corpus contains all the speeches of the Dutch *Tweede Kamer*, where all topics, speeches and interruptions are hierarchically stored in an open XML standard. Each speech or interruption is labeled with the name and parliamentary role of the speaker. The corpus of parliamentary meeting notes is publicly available but awaits a modern, easy-to-use search system which allows users to express complex requests.

As an example of a complex information request, consider a searcher looking for speeches of the Dutch ex-prime minister *Jan-Peter Balkenende* on environmental policies, where he reacts to an interruption by a member of a particular political party. This information request has implicit structural constraints. The returned results must be speeches, not whole documents, and only the speeches where *Jan-Peter Balkenende* is the speaker (in his role as Prime Minister) are relevant, but again only when interrupted by a member of the requested political party. The content of the speeches must be about environmental policies.

State-of-the-art systems are often designed from a data- or system-centered perspective but are suboptimal from a user-centered perspective. They typically allow either free-text retrieval or structured database search. With free-text retrieval, users can type keywords, and documents are returned that contain these keywords, ordered by relevance (which is based on how often these keywords occur). However, they are not suitable for dealing with structural conditions such as restricting speeches to those spoken by *Balkenende* as Prime Minister where he is interrupted by female members of parliament. Databases, on the contrary, can easily deal with these structural conditions, because they capture structural information in fields which can be queried individually, but are not designed to deal with degrees of relevance and interpreting the content of natural language texts. We have developed the retrieval techniques that can deal with an information request like this, and greatly reduce the effort of locating the relevant speeches within these long documents. At the same time, showing aggregated results over the whole debate, or larger units, highlight the actual political processes at stake, and how they evolve over time.

We have shown that (especially expert) users have complex information needs, and that search effectiveness can be improved by using complex queries [10], users tend to have difficulty to use complex query languages to transform their complex information needs in a proper structured query [e.g., 5–7]. Hence there is a need for supporting searchers during their whole task or search episode – both when interactively constructing a complex query, and when interactively exploring the result space [e.g., 3, 4, 9]. These results resonate closely with the approach of De Vries and Spinque, who have developed the Search-by-strategy concept [1, 2] to bridge this gap. This concept separates setting up a search strategy ("the how") from the actual searching and browsing information ("the what"). Search strategies customized for different user profiles are expressed visually, and subsequently transformed automatically into a highly interactive exploratory search interface, where searchers can discover interactively which semantic structures are useful for their particular search problems.

We further increase the value of our system and the political corpus by integrating other resources such as a biographical database (`http://parlement.com/`), a video database (`http://openkamer.tv/`), and generic resources (e.g. `http://dbpedia.org/` and `http://goo.kb.nl/`). These can be added as extra features to exploit structure to give direct access to video material of speeches and biographical data on speakers.

6 Planning

- Planning and phases of activities (7 month period):

  Months 1-2 Pre-processing and indexing of the data. Data (1930-now) is available in preprocessed form (Marx); and can be ingested into the open source MonetDB engine (Spinque); an initial dedicated interface for complex query formulation and interactive result exploration is build (Spinque). Because data and tools are available, this phase can be executed in a short time. Launch of first version (alpha) suitable for initial testing. In parallel, a design tailored to the use cases and target audience is developed (Kamps, Marx).

  Months 2 Expert meeting with developers and users on alpha version and the design and possible extensions.

  Months 3-5 second version (beta) based on updated specs, and final user interface (UI) and user experience (UX) design (Spinque). Redesign search engine for an improved user search experience (Spinque). Based on extensive tests of live alpha version: different users, adjusting specifications. Define appropriate search strategies tailored to the use cases (Kamps, Marx). Define further and more advanced search strategies tailored to the use cases (Kamps, Spinque).

  Months 5 Evaluation meeting with target users on advanced use cases using the beta version, aiming for refining the design and possible extensions.

  Months 6-7 Extensive tests of live beta version: updates and delivery.

  Month 7 Launch and dissemination event (in the KB or Tweede Kamer). Strategic (internal) meeting on future extensions and possible funding opportunities.

- Division of labour:
  - Spinque: responsible for the spiral development process, server back-end creation and tuning, and carry out UI/UX design.
  - University of Amsterdam: project management, responsible for functional requirements, access to all project results, and outreach to prospective user communities.
- Start and end dates: May 1st, 2012–December 1st, 2012.
  - Preparatory work will start in April 1st, 2012.

7 Begroting

We request a total budget of 30,000 Euro.

Cost Type	Activity	Request Funding?	Costs
Personnel	Embedded researcher (7 pm, 0.2 fte)	No	16k
Personnel	Spinque (7pm, 0.15 fte)	Yes	10k
Personnel	Spinque (7pm, 0.15 fte)	No	10k
Material	Expert meeting (month 2)	Yes	3k
Material	Evaluation meeting (month 5)	Yes	3k
Material	Launch/Dissemination event (month 7)	Yes	4k
Material	Domain/Server/Hosting	Yes	10k +
Total			56k
Requested budget			30k

- Estimated 1.4 person-months for researchers (estimated at 16k) will be matched from Kamps' sabbatical leave – he will stay as embedded researcher at Spinque for the total of 1 day a week.
- Spinque estimates 2.1 person-months for the user-centered design and implementation of the operational system: including functional design, database back-end server, UI and UX design and implementation, etc., in a spiral development. Spinque agreed to match the project's contribution so effectively contribute the same amount to the project.
- Both expert meetings about the design (month 2), evaluation meetings with representatives of the prospective user population (month 5), and a final launch/dissemination event (month 7) are scheduled to actively engage all stake-holders.
- A total of 10k for a server plus hosting. The resulting tools can be housed in the University of Amsterdam (either at the center for digital humanities, or at the faculty of humanities) and directly linked from other domains (e.g., the KB site or its DSG portal).

8 Maatschappelijke waarde

Relevance The relevance of this project is threefold: First, the Dutch government wants to make the parliamentary notes openly accessible to make their work transparent for the general public. A search systems that exploit our project results can improve access to this political data, which brings people closer to their chosen government and makes it easier for them to get informed on and participate in the political debates within society (democratic relevance).

Second, the project fits well in the current open government data trend, started by Obama with the data.gov initiative and strongly supported by Sir Tim Berners Lee. We see that more and more government data is becoming available, but that applications which make this data publicly available are often ad-hoc and tailored to a fixed static set. These applications basically tell the public what is in the data. They do not let the public explore the data or create new value with the data. The tool that we propose to build instead gives people this opportunity (social relevance).

Third, Spinque is a Dutch start-up company, and can improve their portfolio with this project, through building a practical application to demonstrate their product (economic relevance).

Target audience Our project will be of interest to several different groups. For the political data, the target audience consists of politicians, civil servants, journalists, historians and political scientists and students who want to want to study certain political topics, politicians and parties or particular events [8], as well as the general public. Commercial companies are interested in the transformation of our knowledge about users and their search behaviour to interface design. A publicly accessible search system based the most recent insights from structured information retrieval has interesting new opportunities for information science research. Information scientists can study real users performing complex search tasks using structured information search systems in a more natural environment.

9 Risico's en afhankelijkheden

Access to Data A risk is the availability of the needed data sets. This risk is very low, since all data is in the public domain, and the pipeline to harvest and preprocess the data is set up in earlier projects.

Information Access Tools A risk is the availability of the required search technology. This risk is very low, since in related projects the required structured search technology has been developed and proof-of-concept implementation have been built successfully.

System efficiency and robustness A risk is the robustness and efficiency of the resulting system. This risk is low and has been addressed by teaming up with a commercial partner that has extensive experience in this.

Relevance to target audience A risk is that the resulting applications don't match the types of tasks of the targeted user population. This risk is low and has been addressed by involving a wide group of representative users in the design and evaluation of the tools. In addition dissemination events will help raising awareness.

10 Haalbaarheid

The project setup and planning is geared toward minimizing risks and ensuring the project's goals are realized.

11 Organisatie

The project is a result of long-standing interactions and collaborations between researchers working on novel retrieval methods for structured data, stakeholders from industry and the cultural heritage domain, and representatives of the prospective user groups.

Name	Title	Role	Expertise	Affiliation
H. Huurdeman	M.Sc./M.A.	Adviser	UI/UX	University of Amsterdam
J. Kamps	Dr.ir.	PI	CH retrieval	University of Amsterdam
M. Koolen	Dr.	Adviser	CH retrieval	University of Amsterdam
M. Marx	Dr.	Co-applicant	Political science & XML	University of Amsterdam
A.P. de Vries	Prof.dr.ir.	Co-applicant	Databases, IR, and multimedia	CWI & TU Delft
W. Alink	Ir.	Commercial	Search by Strategy, UI/UX dev.	Spinque B.V.
P. Doorenbosch	Drs.	Heritage	*Staten Generaal Digitaal*	Koninklijke Bibliotheek
J. Keukens	Drs.	User group	*Dienst Informatievoorziening*	Tweede Kamer
J. van de Merrienboer	Dr.	User group	*Centrum Parlementaire Geschiedenis*	Radboud University
Th. Verkade	Drs.	User group	*Afdeling data-journalistiek*	NRC
M. Visser	Drs.	User group	*Redactie internet en Den Haag*	Trouw
R. Vliegenthart	Dr.	User group	ASCOR	University of Amsterdam

Hugo Huurdeman (Media Studies, University of Amsterdam) is the Ph.D.-candidate in the WebART project funded by NWO/CATCH, and previously owner of `http://www.timelessfuture.com/` developing Web-based multimedia solutions based on advanced user interface/user experience design.

Jaap Kamps (Archives and Information Studies, University of Amsterdam) is an expert on Information Science and Information Retrieval with over 250 publications.[1] He is the leading chair of the INEX evaluation forum with an annual cycle studying all aspects of focused retrieval on structured text.[2] He has organized a dozen workshops and conferences, and is teaching a range of courses on modern information access within the Faculty of Humanities. He is the PI of four large research projects that received funding from NWO: the *Multiple-collection searching using metadata* (MuSeUM) project funded by the Continuous Access to Cultural Heritage (CATCH) program; the *Effective Focused Retrieval Techniques* (EfFoRT) project funded by the *Vrije Competitie*; the *Retrieving Encoded Archival Descriptions More Effectively* (README) project funded by the Innovative Research Grants Program (VIDI Scheme); and the *Web Archive Retrieval Tools* (WebART) project funded by the Continuous Access to Cultural Heritage (CATCH) program; In addition, he is involved as a co-applicant in two European and two Dutch funded projects.

Marijn Koolen (Archives and Information Studies, Humanities, University of Amsterdam) is currently post-doctoral researcher in the README project. He graduated from the University of Amsterdam while working on a NWO/CATCH project on unified access to a Museum's register, library catalogue, and exposition archive.

Maarten Marx (Informatics Institute, University of Amsterdam) is a leading researcher on querying highly structured textual data, especially when cast in XML or its variants, and the leader of the `http://politicalmashup.nl/` project. Maarten will make the parliamentary proceedings available in an enriched XML format, and contribute a number of use-cases.

Arjen de Vries (TU Delft & CWI Amsterdam) is Professor of Multimedia Spaces at Technical University Delft, and senior researcher at the national research institute for mathematics and computer science (CWI). He is a renowned database and information retrieval expert, and multimedia retrieval specialist. Arjen will function as a technical advisor.

[1] See `http://staff.science.uva.nl/~kamps/publications/`.

[2] See `https://inex.mmci.uni-saarland.de/`.

The industrial partner in the project is `http://spinque.com/`, and lead developer Wouter Alink. Spinque is a recent start-up company specializing comprehensive search on structured data. We have already operational prototypes within the running projects. Spinque can turn these into a fully-fledged operational service, using their experience in the design and optimization of high volume web-sites. Prof. De Vries is co-founder of Spinque, ensuring seamless communication and transfer of knowledge between the proposed project and the industrial party. Moreover, Dr. Kamps will be an embedded researcher at Spinque, as part of his sabbatical year in 2012.

The heritage partner is the National Library of the Netherlands. Paul Doorenbosch (Koninklijke Bibliotheek) is head of the Research Department at the National Library of the Netherlands and is involved in several collaborations between Computer Science research, cultural heritage and humanities. He also leads the `http://statengeneraaldigitaal.nl/` project.

The project is backed up by a range of potential users of the parliamentary proceedings: Jan Keukens (Tweede Kamer) is senior member of the Department of Information Services, House of Representatives of the *States General*. Johan van de Merrienboer (Centrum voor Parlementaire Geschiedenis, Radboud Universiteit Nijmegen) studied History and Dutch law in Utrecht, received his Ph.D. in 2006, and is a researcher at the *Centre for Parliamentary History* at the Radboud Univeristy Nijmegen. Thalia Verkade (NRC) is a journalist of the nrc.next, specializing on sense-making of large data-sets in particular those of the government. Marco Visser (redactie internet en Den Haag, Trouw) is online journalist for Trouw, writing on politics. Rens Vliegenthart (University of Amsterdam) is an assistant professor at the Amsterdam School of Communication Research, studying the interaction between politics and media. The user group will play a crucial role during the design of the innovative search tools tailored to complex (re)search requests that cannot be answered using traditional access methods.

12 Relatie met bestaand onderzoeksprogramma?

The project stems from the combination of recent insights in novel ways of accessing structured information (Kamps) to the richly structured and annotated data of the parliamentary proceedings (Marx).

- Lead project members (Kamps, Koolen, Huurdeman) provide insight in modern access methods to structured data as is emerging in a range of related projects (such as MuSeUM, README, EfFoRT, WebART). These insights will be applied to a new and important genre of data – parliamentary proceedings – which is of clear importance in its own right, but also a representative of other meeting notes and transcripts that constitute a large part of governmental or archival data.
- Over the last decade, Kamps has been working closely with partners Marx and De Vries on access to structured data.
  Prior work with Marx on political data will be extended. The NWO funded PoliticalMashup project will provide Dutch Hansard data in XML format and linked to a Biographical database (`http://parlement.com/`) and a video database (`http://openkamer.tv/`). At the time of writing all debates from Parliament from 1930 to November 2010 are available.
  Prior work with De Vries on supporting complex search tasks will be extended by casting them as a "Search by Strategy" approach. This will mainly be done through our collaboration with Spinque.
- Commercial partner is Spinque, a company specializing in technology that exploits structural information for search systems and in the valorisation of innovative research results on structured information retrieval. With their search-by-strategy concept and expertise in designing structured information retrieval systems, they can turn existing prototypes into proper industry-strength online services, significantly improving access to this valuable resources for hundreds or thousands of interested users.
- Heritage partner the National Library of the Netherlands (Doorenbosch), e.g., `http://statengeneraaldigitaal.nl/`, have a vested interest in novel access methods for the parliamentary proceedings.
- User groups are the *Dienst Informatievoorziening Tweede Kamer*, and journalists, political science students, communication science students, general public. They will be instrumental in providing insight into the types of information requests they have and how these can be transformed into effective and intuitive search strategies.

13 Deliverables/Concreet eindproduct

The deliverables of the project are:

- Expert meeting on exploratory political search, and a deliverable report on the meeting.
- Evaluation meeting (living lab style) where all stakeholder together work on specific use cases of research made possible by the data and search tools available, and a deliverable report on the meeting.

- A fully functional search system that gives access to a growing corpus of structured political documents, that exploits the available structure to allow and help users to express their complex information needs, and explore the resulting space through a combination of facetted search and visual analytics – all based on a generic underlying data model.
- A launch/dissemination event – attracting wider attention in the press – and a report on this event.
- Strategic report on the next challenges and solutions and possibilities to obtain further funding to advance this new research direction (internal). Initial ideas are on: i) further expanding the current focus; on ii) completing the parliamentary history by including earlier archives of the *Staten Generaal* (1576–1796) and the *Bataafse Republiek* (1796–1813) (*Nationaal Archief* and Huygens/ING); and on iii) deploying the techniques on a wider range of data (contacts with KB and `http://stichting.bibliotheek.nl/`).

14 Techniek

All parliamentary proceedings are publicly available at the National Library of the Netherlands (1814–1995, `http://statengeneraaldigitaal.nl/`) and the Dutch parliament (1995–now, `http://overheid.nl/`). Basic search interfaces are available at both sites.

A pipeline to automatically harvest the proceedings and render them into a rich XML format is constructed as part of Dr. Marx's (`http://politicalmashup.nl/`). The rich format brings out the debate structure (who said what to wham and in what context) and provides rich background on the speakers (e.g., party membership but also various other demographics), and the topic of the debate.

Baseline architecture will be based on the open source MonetDB engine (`http://monetdb.org/`). The resulting engine is very powerful but requires programming skills far beyond what can be expected from the target audience. Spinque has a proprietary editor that can be used for fast prototyping of dedicated advanced search strategies [e.g., 3, 4, 9]. The resulting search strategies and their implementation will be released as open source to the community.

A dedicated server will be set up, likely within the UvA, but resolvers are available that can make the services available from the `http://kb.nl/` domain. Placement within the UvA domain will allow direct and intensive access for local scholars. An attractive option is to integrate this server in a shared facility of the Center for Digital Humanities – since adequate ICT support is a shared concern of many computational intensive projects in the Faculty of Humanities.

Literatuurverwijzingen

[1] A. de Vries, W. Alink, and R. Cornacchia. Search by Strategy. In J. Kamps, J. Karlgren, and R. Schenkel, editors, *ESAIR '10: Proceedings of the third workshop on Exploiting semantic annotations in information retrieval*, pages 27–28, New York, NY, USA, 2010. ACM.

[2] A. P. de Vries. What to do when one size does not fit all? In O. Alonso, J. Kamps, and J. Karlgren, editors, *ESAIR'11: Proceedings of the fourth workshop on Exploiting semantic annotations in information retrieval*, pages 1–2, New York, NY, USA, 2011. ACM.

[3] J. Kamps. The impact of author ranking in a library catalogue. In G. Kazai, C. Eickhoff, and P. Brusilovsky, editors, *Proceedings of the Fourth Workshop on BooksOnline'11: Online Books, Complementary Social Media, and Crowdsourcing*, pages 35–40. ACM Press, New York NY, 2011.

[4] J. Kamps. Toward a model of interaction for complex search tasks. In O. Alonso, J. Kamps, and J. Karlgren, editors, *Proceedings of the Fourth Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR 2011)*, pages 7–8. ACM Press, New York NY, 2011.

[5] J. Kamps, M. Marx, M. de Rijke, and B. Sigurbjörnsson. Structured queries in XML retrieval. In A. Chowdhury, N. Fuhr, M. Ronthaler, and H.-J. Schek, editors, *CIKM'05: Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 2–11. ACM Press, New York NY, USA, 2005.

[6] J. Kamps, M. Marx, M. de Rijke, and B. Sigurbjörnsson. Understanding content-and-structure. In A. Trotman, M. Lalmas, and N. Fuhr, editors, *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology*, pages 14–21. University of Otago, Dunedin New Zealand, 2005.

[7] J. Kamps, M. Marx, M. de Rijke, and B. Sigurbjörnsson. Articulating information needs in XML query languages. *Transactions on Information Systems*, 24:407–436, 2006.

[8] R. Kaptein and M. Marx. Focused retrieval and result aggregation with political data. *Information Retrieval*, 2010.

[9] R. Kaptein, M. Marx, and J. Kamps. Who said what to whom? capturing the structure of debates. In J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, and J. Zobel, editors, *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 831–832. ACM Press, New York NY, USA, 2009.

[10] M. Koolen and J. Kamps. Searching cultural heritage data: Does structure help expert searchers? In *Proceedings of RIAO 2010: Adaption, personalization and fusion of heterogeneous information*. C.I.D. Paris, France, 2010.