

PoliticalMashup

Large scale integration of political data

Maarten Marx

Universiteit van Amsterdam

IvI Colloquium, Science Park Amsterdam, 2011-05-24



Goal of this project

Enabling Computational Social Science on Political Data.

- Diachronic

Goal of this project

Enabling Computational Social Science on Political Data.

- Diachronic
- Compare political actors

Goal of this project

Enabling **Computational Social Science** on Political Data.

- Diachronic
- Compare political actors
 - ★ Parties
 - ★ Politicians
 - ★ Roles
 - ★ Executive Organizations, like
 - nation states
 - ministeries



Example: competing for attention

Economy of attention

[Huberman et al, CACM 2010]

- **Idea:** attention = being interrupted during your speech.

Example: competing for attention

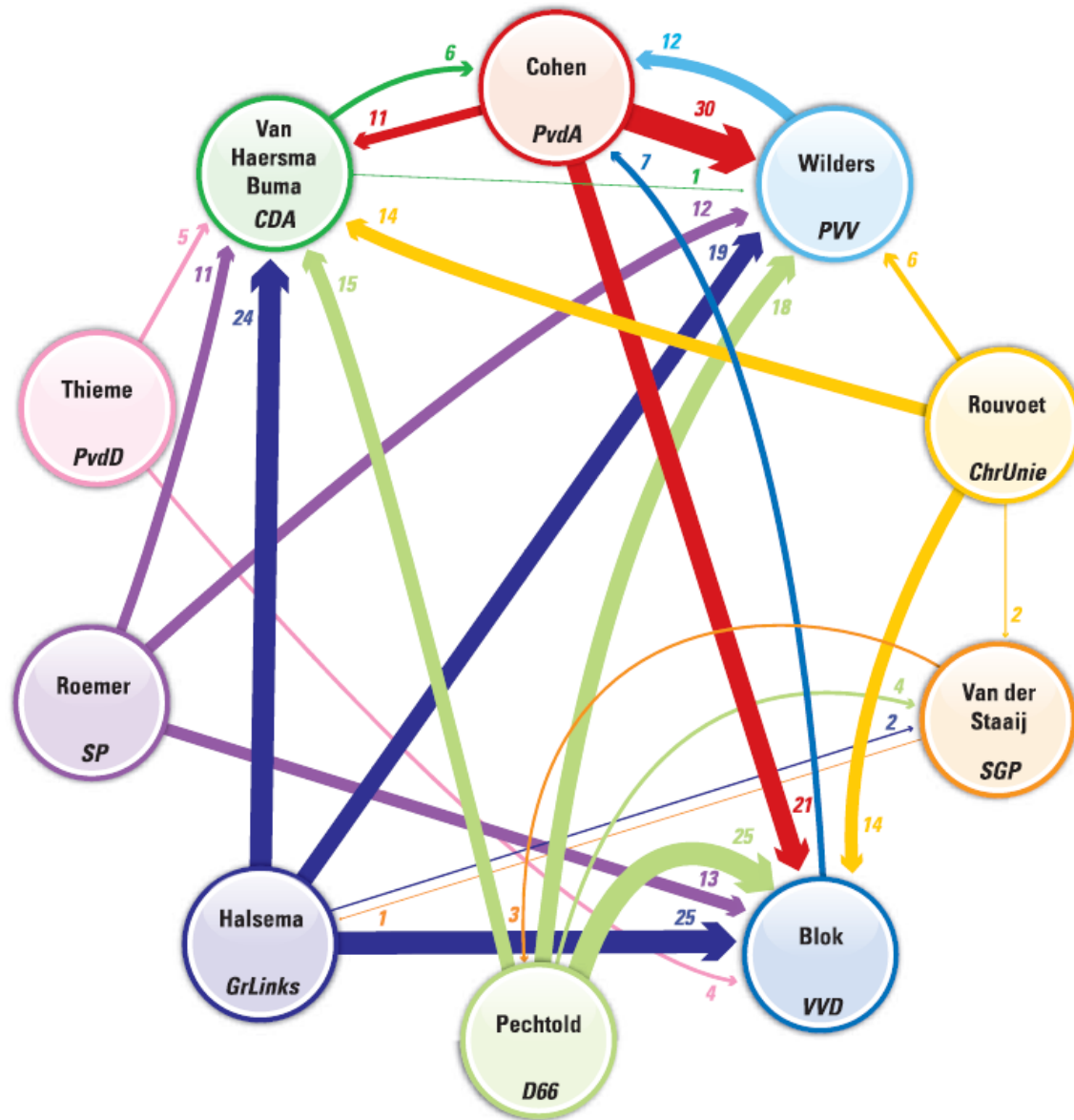
Economy of attention

[Huberman et al, CACM 2010]

- **Idea:** attention = being interrupted during your speech.
- Each person gives and receives attention to and from other persons.
- Each event is a directed network of persons giving and receiving attention.



Attention network

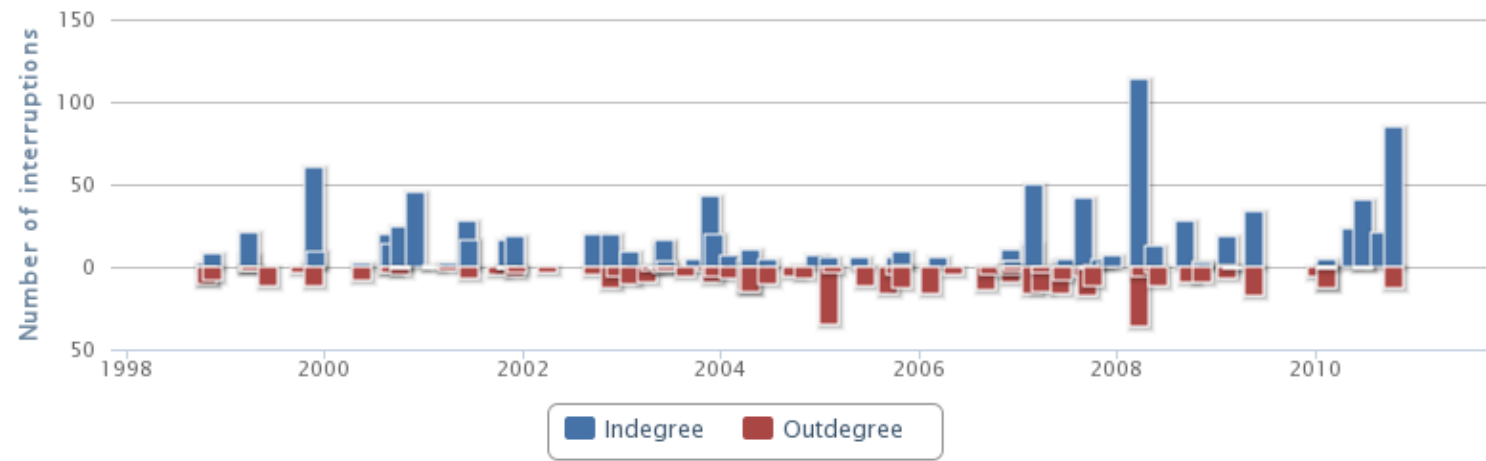
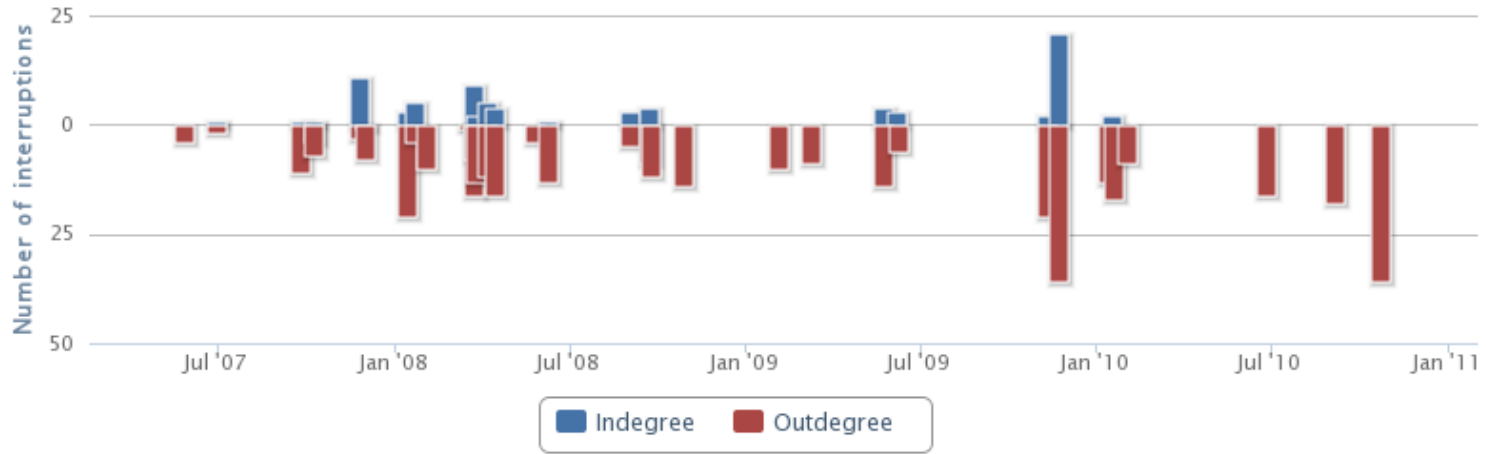


Diachronic view on one person

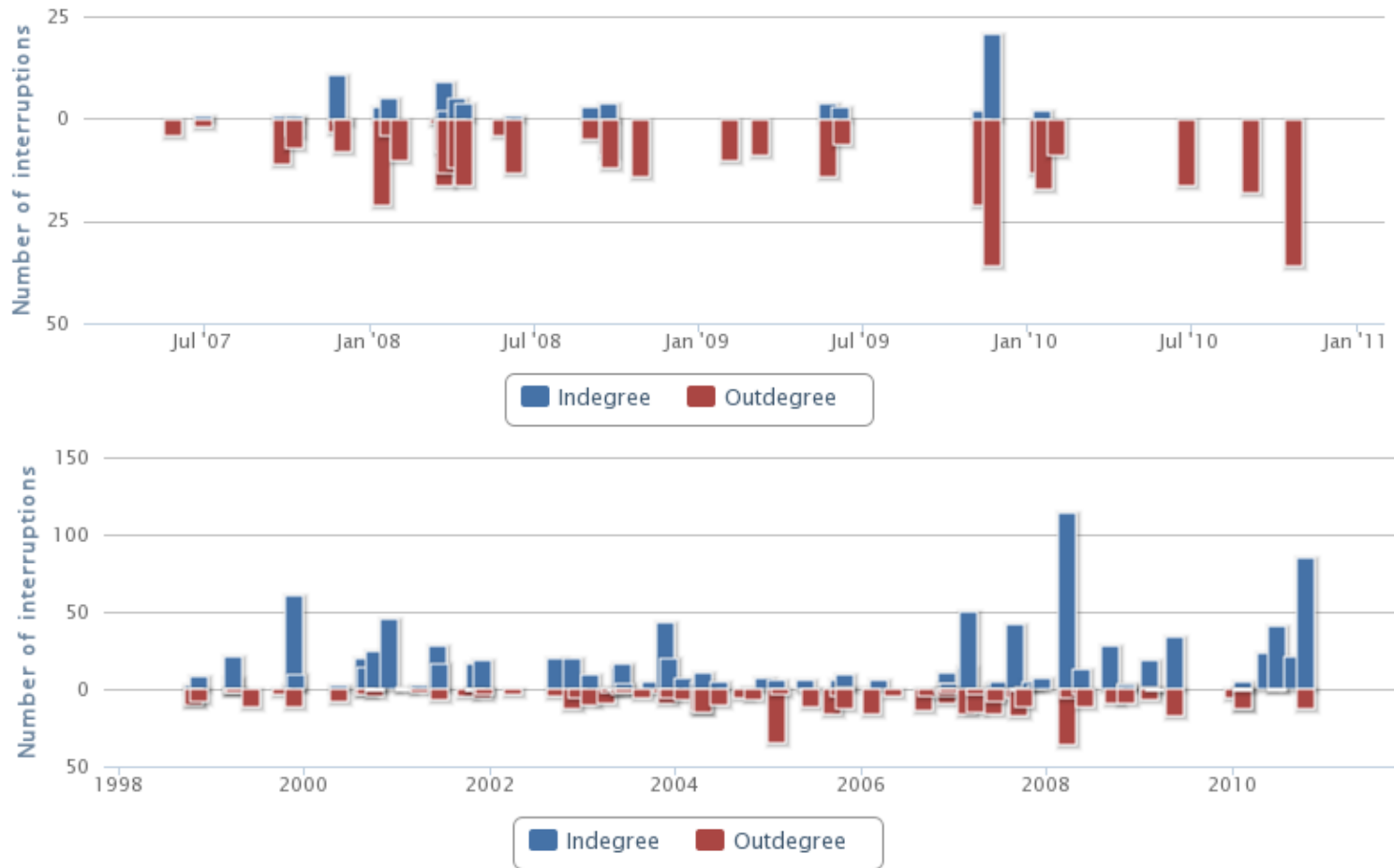
<http://xml.politicalmashup.nl/XQueries/debates/inoutdegree.xq?mpid=02207>



Compare careers of different politicians



Compare careers of different politicians



Interruption in- and out-degree of Roemers and Wilders.

Important dimensions for comparison of data

- Time
- Political Actors
- Issues
 - ★ E.g., compare **PVV** and **SP** on **immigration**-topic through time.

Requirement:

- all useful data units are **uniformly** tagged with **normalized** values on these dimensions.



Data upside down, or inside out

- Often, scientific analysis asks for a radically different organization of the data.
- Take DBLP as an example.

dblp .uni-trier.de
Computer Science
Bibliography



From a document centric view

Het internet Afbeeldingen Video's Maps Nieuws Shopping Gmail meer ▾

Google scholar Peter Sloot Zoeken [Geavanceerd zoeken met Scholar](#)

Het internet doorzoeken Zoeken in pagina's in het Nederlands

Scholar elke datum inclusief citaten [E-mailmelding maken](#)

Tip: [alleen in het Nederlands zoeken](#). U kunt uw zoektaal bepalen in [Voorkeuren voor Scholar](#).

[The distributed ASCI supercomputer project](#) [\[PDF\] van psu.edu](#)
 UvA-linker: Full Text
 ..., B Overeinder, P Sloot... - ACM SIGOPS ..., 2000 - portal.acm.org
 ... Hendrikse, Bob Hertzberger, Alfons Hoekstra, Kamil Iskra, Drona Kandhai, Dennis Koelma, Frank van der Linden, Benno Overeinder, Peter Sloot, Piero Spinnato
 Department of Computer Science, University of Amsterdam Dick ...
[Geciteerd door 95](#) - [Verwante artikelen](#) - [Alle 23 versies](#)

[Lattice-Boltzmann hydrodynamics on parallel systems](#) [UvA-linker: Full Text](#)
 ..., M Kataja, J Timonen, PMA Sloot - Computer Physics ..., 1998 - Elsevier
 Realistic lattice-Boltzmann simulations often require large amounts of computational resources and are therefore executed on parallel systems. Generally, parallelization is based on one- and two-dimensional decomposition of the computational grid in equal subvolumes, and load ...
[Geciteerd door 87](#) - [Verwante artikelen](#) - [Alle 7 versies](#)

[Effect of nutrient diffusion and flow on coral morphology](#) [\[PDF\] van uu.nl](#)
 UvA-linker: Full Text
 ..., CP Lowe, D Frenkel, PMA Sloot - Physical review letters, 1996 - APS
 We describe a method for modeling aggregation in a flowing fluid. In the model, aggregation proceeds by the accumulation of a "nutrient." The nutrient is modeled using a lattice Boltzmann model of transport. The aggregate absorbs the nutrient, and the amount absorbed ...
[Geciteerd door 76](#) - [Verwante artikelen](#) - [Alle 15 versies](#)

[Towards a grid management system for HLA-based interactive simulations](#) [UvA-linker: Full Text](#)
 ..., M Bubak, M Malawski, P Sloot - 2003 - computer.org
 This paper presents the design of a system that supports execution of HLA distributed interactive simulations in an unreliable Grid environment. The design of the architecture is based on the OGSA concept that allows for modularity and compatibility with Grid Services already ...
[Geciteerd door 66](#) - [Verwante artikelen](#) - [Alle 9 versies](#)

[Mesoscopic simulations of systolic flow in the human abdominal aorta](#) [\[PDF\] van psu.edu](#)
 UvA-linker: Full Text
 ..., AG Hoekstra, PMA Sloot - Journal of Biomechanics, 2006 - Elsevier
 The complex nature of blood flow in the human arterial system is still gaining more attention, as it has become clear that cardiovascular diseases localize in regions of complex geometry and complex flow fields. In this article, we demonstrate that the lattice Boltzmann method ...
[Geciteerd door 64](#) - [Verwante artikelen](#) - [Alle 14 versies](#)

[From molecule to man: Decision support in individualized e-health](#) [\[PDF\] van psu.edu](#)
 UvA-linker: Full Text
 PMA Sloot, A Tirado-Ramos, I Altintas... - Computer, 2006 - ieeexplore.ieee.org
 PUSHING AND PULLING An application pull has occurred in biomedicine with the move to in silico studies, which augment in vivo and in vitro studies by simulating more details of biomedical processes. Using these simulated processes helps medical doctors make ...
[Geciteerd door 54](#) - [Verwante artikelen](#) - [Alle 13 versies](#)



To an actor centric view

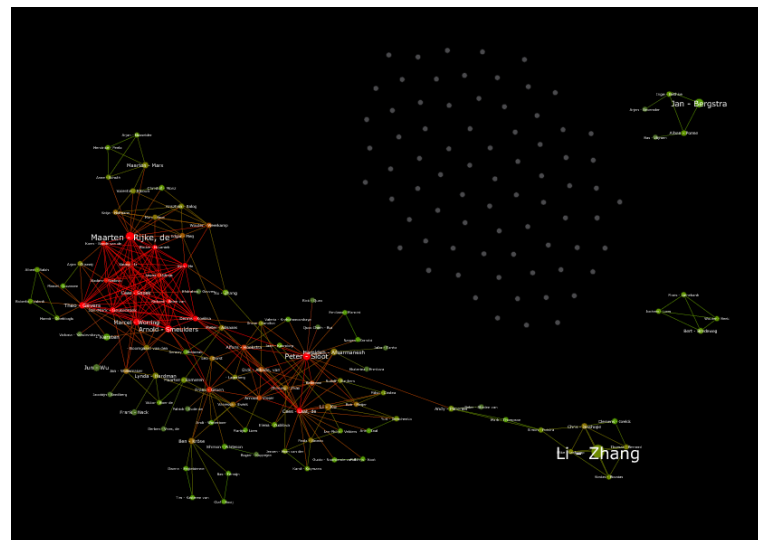
Compare

- Current Ivl Director: <http://dblp.uni-trier.de/search/author?author=Deniz+van+Heijnsbergen>
- Previous Ivl Director: <http://dblp.uni-trier.de/search/author?author=Peter+Sloot>



New view creates new opportunities

- “Easy” to create a cooperation graph between Ivl members.
- Detect existing clusters of people, persons which span different groups, ...



Creating DBLP sounds easy....

- Just collect all issues of all journals
- Extract the author-article-title (bibtex) information
- Normalize author names, dates, journal names, ...
- ...
- keep it up to date



But it is not

when you want to be and remain

- complete
- correct
- up to date

Compare what you need to do to collect your own data for the annual report ...

Main Problems:

- scale
- continuous change
- interdependencies



Back to Political data

Similar to DBLP, but we even want more ...

- from document- to person-centric view
- plus extract structure and content from documents
- plus classify/tag all relevant parts of documents
- also on (really old) legacy data.



Example Structure Extraction

- Rich data model [Link to example](#)
- Meeting (1 Day)
 - Topic
 - Stage direction
 - Scene
 - Stage direction
 - Speech
 - Paragraph



Same data: different views

- Raw data in PDF
- human readable XML
- Machine readable XML



Applications on legacy data

- Search for documents in which Anne Vondeling speaks.
- Connect named entities to biographical database.
- Connect roles (Chair, Minister of Justice, etc) to persons using biographical data.
- Main challenge: OCR errors

BIJKEKHOEPINGSUK IJ VOOHMIDDAG.

([Link naar Handeling](#))



Some technical details



Extract-Transfer-Load [Rahm Do 2000]

- Step 1a** Data cleanup (encoding issues, OCR-error correction with TICCL [Reynaert 08])
- Step 1b** (Optional) Basic structure recognition (with hand crafted rules) **OUTPUT: XML**
- Step 2** Specify data transformation **declaratively** in XPath (= First Order Logic of Trees)
- Step 3** Compile to XSLT transformer and transform



Extract-Transfer-Load (continued)

Step 4 Validate data with

- regular tree automata (Relax NG)
- integrity constraints and dependencies (Schematron)

Step 5 Load into combined XML database and search engine (eXist/Lucene).



Learning Classifiers

- Usual Problem
 - little trainings data
 - trainings data on wrong aggregation level
- Luck for election manifestos
 - Isaac Lipschits manually annotated **every paragraph** in the election manifestos between 1977 and 1998 with **controlled vocabulary terms**.
 - **Surjective** back of the book index.
 - Strongly hyperlinked document. [Link to 1998 Volume](#)
 - We have scanned these books, will turn them into a hyperlinked database, and train classifiers.



TODO: investigate temporal robustness of classifier

Summing up

- We turn a collection of text documents into a database.
- Nothing manual.
- Make implicit structure explicit. Classify small textual units.
- Allowing social scientists to perform **traditional DB and IR tasks** on semi-structured data
 - Search with complex content and structure queries
 - Analysis/ Data Cubes
 - Data quality management (integrity, consistency)
 - Define Views.



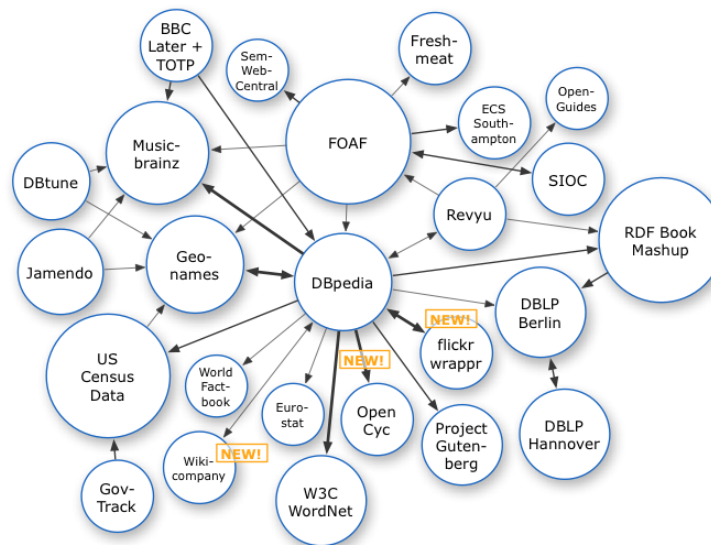
Why do all this? What are the gains?



Long term

Become a core component of the [linked open government data cloud](#) [Berners Lee 2009].

- Compare with DBLP.
- Since the beginning a core component of the LOD.



Medium and short term

This very rich data source works well in **connecting** Computer Science to

- Foundational CS research
- Humanities and the Social Sciences
- Society



Testbed and Inspiration for Foundational Research

- FP7 STREP Foundations of XML project
 - Benchmark and test data
 - Document centric data exchange
Neven et al, PODS 2011
NWO EW Open Competition
 - Faceted Search Evaluation [INEX 2011]



Computer Science applied in other sciences

- [UvA ASCoR](#) (Humanities) : compare populist parties in last 20 years in NL, Belgium, Denmark, Sweden (CCCT funded).
- [NIOD](#) (Humanities): [War in Parliament](#) project (Clarin funded).
- [DNPP](#) (Political Science): web and media use of parties and politicians. Agenda setting research (NWO funded).



Benefits for society

- Improved information access for general public and professionals through
 1. Cooperation with Dutch Parliament
 2. Statengeneraaldigitaal.nl (Koninklijke Bibliotheek)
 3. Research → Media → Kamervragen → Serious pressure for improvement.

Op welke wijze gaat u zich inzetten voor volledige, juiste en digitaal makkelijk toegankelijke informatie uit het parlement?

[P. Heijnen, PvDA 2011-03-04]



End

Thanks to

