DIGGING INTO PARLIAMENTARY DATA

JAAP KAMPS AND MAARTEN MARX

Abstract. This is a tutorial proposal for DL'14. The tutorial will give attendees insight in the added value of semantically enriched textual corpora, focusing in particular on cases where the annotation reflects the data and use cases. This is demonstrated on parliamentary proceedings, a relatively generic genre of text transcripts characterized by internal structure and associated entities such as speakers/writers/.... We showcase the remarkable power of the semantic annotation of debate structure and speakers through numerous examples that demonstrate how linear text becomes valuable research data that can be sliced and diced to present many views. We also give a look under the hood and learn attendees about the relative ease of incrementally annotating text, starting from OCR'ed flat text to encoding structure and entities, in ways that connect to other information sources (such biographies of speakers in Wikipedia/DBpedia).

1. Tutorial Content

1.1. **Tutorial Abstract.** The proceedings of parliament are a key example of textual data which contain rich metadata about *every word in the text.* For every word spoken in the meetings of parliaments, the proceedings record who said it, in which role, on behalf of which party, when it was said, in the context of which debate, and sometimes even to whom it was said. In many countries, these proceedings have recorded for centuries in a more or less constant and consistently applied manner, providing a great longitudinal corpus connecting the history to the present, and to the future. Proceedings are thus an ideal source for diachronic comparative research. Recent digitization efforts made large data sets available (e.g, in the UK since 1835, in the Netherlands since 1814), but most often only as OCR'ed scans with only scarce (because manual and thus expensive) metadata at the document level.

This tutorial shows how one can enlarge the value of such data sets by making the implicit, human readable, structure explicit in machine readable XML markup. We cover existing schemas and standards, best practices, and examples; we show a number of example research questions, ranging from data journalism to historical scientific research, which can be answered using the marked up data; and in an hands on setting we show how relatively easy it is to transform flat text into deeply annotated data. We conclude by taking a larger perspective, going beyond parliamentary proceedings, indicating that the steps taken in this tutorial also apply to many other data set in the (digital) humanities.

1.2. **Topical Outline of the Tutorial.** The tutorial consists of three parts.

The first part is about the nature of parliamentary proceedings, what valuable (implicit) structure is located in such documents, how we should represent that using mark up languages and/or using Linked Data (RDF), and what kind of metadata standards are available for such data. This part is organized as a running example of the pipeline starting from raw text (*what* is said) to enriched data (*who* it says, and *to whom* it is said, and *why*).

In the second part we show what this enables. So what new and powerful ways of viewing and analysing the data become available by transforming text with implicit structure into heavily marked up documents in which the structure is made explicit. We do so by going through a number of exciting examples, ranging from websites like `http://TheyWorkForYou.com/` to creating profiles of politicians and parties based on speeches.

The aim of the third part is to look under the hood and show that the data-transformation process is often much easier than one may think, and can be done in an incremental manner. All transformations are automatic and scale to large data sets, relatively straightforward ad hoc methods exploiting textual cues and the implicit structure are remarkably effective. This part is hands-on: we use XPath and XSLT to transform OCR'ed scans into richly marked up XML.

The tutorial material will be made available before the tutorial, including access to the data and tools discussed in the tutorial plus links to larger data sets. We actively encourage tutorial attendees to further explore the examples, and their own variants, before, during, and after the tutorial.

2. Tutorial Facts

2.1. **Duration.** We prefer a half-day tutorial.

2.2. **Expected Number of Participants.** We estimate some 20-25 participants.

2.3. **Target Audience and Level of Experience.** The tutorial is of interest to a large fraction of DL attendees. It is of interest to researchers and practitioners in digital libraries, in particular those working with or interested in domain-specific collections and like to learn of state of the art work on semantic annotation and its use. It is also of interest to people working with large historical text archives which have a consistently applied structure. These can be digital humanities scholars, librarians, archivists, documentalists, but also historians who want to apply digital methods to large data sets.

The level is introductory. We assume some general familiarity with mark-up languages and metadata standards. We will get a bit technical when discussing text transformations, but all parts will be illustrated by real working examples, and DIY instructions (allowing for attendees to put what they learn into practice).

2.4. **Learning Objectives.** After the tutorial you have a clear idea of the ideal format in which to store and supply a digital parliamentary data set. You understand the value for diachronic comparative research of parliamentary data structured according to standard schemas and metadata standards. You have seen a number of exciting examples of such research. And you have seen that turning a purely textual set of parliamentary proceedings (e.g. OCR'ed scans) into heavily marked up XML is much easier than you may think.

2.5. **Biographical Sketch.** Dr. Jaap Kamps is Associate professor of Information Retrieval at Archives and Information Studies, Department of Media Studies, Faculty of Humanities, University of Amsterdam. His research interests are information storage and retrieval, big data, linked data, structure and semantic annotation, digital humanities, e-humanities, digital heritage, novel access methods for digital information, evaluation and user studies, interactive search, task based search, exploratory search, sense making. He is PI of a range of externally funded research projects on the search and exploration of domain specific collections from libraries, archives, and museums. He is an active organizer in DL and IR conferences and workshops, in particular focusing on richly annotated corpora (e.g., INEX, CLEF, TREC, ESAIR series).

Dr. Maarten Marx is Associate Professor at the Informatics Institute, Faculty of Sciences, University of Amsterdam. His research interests combines theoretical research in XML database and search technology with applications of that research in eHumanities projects involving transforming and integrating large amounts of mostly textual data. He advises the Internal Information Service (Dienst Informatie Voorziening) of the Dutch Parliament about modeling the Parliamentary Proceedings in a digitally sustainable manner and about interlinking them with external datasources. His work on the parliamentary proceedings was recognized with the XML Holland Award 2008 and the Dutch Data prize (awarded by DANS-KNAW) 2012.

Kamps and Marx have collaborated closely for 20 years, recent joint projects of relevance to the proposal include *Exploratory Political Search* (ExPoSe, funded by the Netherlands Organization for Scientific Research, `http://www.nwo.nl/en/research-and-results/research-projects/78/2300179278.html`) and *Digging Into Linked Parliamentary Data* (DILiPaD, funded by the Digging into Data Challenge, `http://dilipad.projects.history.ac.uk/`).

2.6. **Related Previous Tutorials.** Not applicable. We build on material from many talks and papers, but the tutorial itself is newly constructed for the DL'14 conference.

2.7. **Contact Information.** Email addresses are `kamps@uva.nl` and `maartenmarx@uva.nl`. Full contact details are on our home-pages: `http://staff.science.uva.nl/~kamps/` and `http://staff.science.uva.nl/~marx/`.